Spring 2014

The treatment of missing data when estimating student growth with pre-post educational accountability data

Jason P. Kopp
James Madison University

The Treatment of Missing Data when Estimating Student Growth with Pre-Post Educational Accountability Data

Jason P. Kopp

A dissertation submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

Department of Graduate Psychology

May 2014

Acknowledgements

I would first like to thank my dissertation and academic advisor, Dr. Sara Finney. You have dedicated an enormous amount of time to providing academic and professional guidance and support throughout my master's and doctoral studies. On this dissertation alone, you have given countless hours to providing feedback on my ideas, reading and critiquing drafts, and have generally supported me throughout this process. This does not include all the time you generously gave nurturing my professional development and helping to guide my future career in the field. I would not be where I am today without your support, and I am very glad that I have had you as an advisor and mentor through these years.

Second, I would like to thank my dissertation committee, Dr. Deborah Bandalos, Dr. Donna Sundre, and Dr. Craig Enders. Thank you for all of your feedback, suggestions, and support throughout the dissertation process.

Third, I would like to thank the entire faculty at the Center for Assessment and Research Studies and in the Assessment and Measurement program. This dissertation would not have been possible without the phenomenal training and support I have received from all of you.

Fourth, I would like to thank all my fellow students in the Assessment and Measurement program and those that have worked in the Center for Assessment and Research Studies through the years. Whether it's late night study sessions, trips to Greenberry's, or unwinding at the Dodger, going through graduate school was all the more enriching by having great friends to share it with. Through all the challenges, thank you for being there to support me.

Finally, I want to thank my family for supporting me throughout this process. I love you all - thank you for all of your love and encouragement throughout my life.

Table of Contents

List of Tables

List of Figures

Abstract

To ensure program quality and meet accountability mandates, it is becoming increasingly important for educational institutions to show "value-added" for attending students. Value-added is often evidenced by some form of pre-post assessment, where a change in scores on a construct of interest is considered indicative of student growth. Although missing data is a common problem for these pre-post designs, missingness is rarely addressed and cases with missing data are often listwise deleted. The current study examined the mechanism underlying, and bias resulting from, missingness due to posttest nonattendance in a higher-education accountability testing context. Although data were missing for some students due to posttest nonattendance, these initially missing data were subsequently collected via makeup testing sessions, thus allowing for the empirical examination of the mechanism underlying the missingness and the biasing effects of the missingness. Parameter estimates and standard errors were compared between the "complete" (i.e., including makeup) data and a number of different missing data techniques. These comparisons were completed across varying percentages of missingness and across noncognitive (i.e., developmental) and cognitive (i.e., knowledge-based) measures. For both noncognitive and cognitive measures, posttest data was found to be missing-not-at-random (MNAR), indicating that bias should occur when utilizing any missing data handling technique. As expected, the inclusion of auxiliary variables (i.e., variables related to missingness, the variable with missing values, or both) *decreased* the conditional relationship between the posttest noncognitive measure scores and posttest attendance (i.e., missingness); however, it *increased* the conditional relationship between posttest cognitive measure scores and posttest attendance. Thus,

utilizing advanced missing data handling with auxiliary variables resulted in reduced parameter bias and reduced standard error inflation for the noncognitive measure, but increased parameter bias for some parameters (posttest mean and pre-post mean change) for the cognitive measure. These effects became more exaggerated as missingness percentages increased. With respect to future research, additional examination of bias-inducing effects when employing missing data techniques is needed. With respect to testing practice, assessment practitioners are advised to avoid missingness if possible through well-designed assessment methods, and to attempt to thoroughly understand the missingness mechanism when missingness is unavoidable.

CHAPTER ONE

Introduction

"Student achievement, which is inextricably connected to institutional success, must be measured by institutions on a 'value-added' basis that takes into account students' academic baseline when assessing their results." (U.S. Department of Education, 2006, p. 4).

It is becoming increasingly important for institutions of higher education to demonstrate the value for students in attending their institution. The cost of college has skyrocketed in recent decades. For example, the total inflation-adjusted cost of a four-year, American public university degree has increased by over 250% since 1982 (College Board, 2012). Despite this increased cost, there is concern among policy makers that students are not receiving adequate education for the dollars they spend (U.S. Department of Education, 2006). Thus, accreditation agencies and other policy makers have demanded tangible evidence of the "value-added" to students attending a given institution. These institutions often attempt to demonstrate value-added by providing evidence of *student growth* over the course of the college career. Student growth can encompass positive changes in cognitive skills (e.g., improved scientific reasoning) or noncognitive traits (e.g., more constructive attitudes towards learning). To adequately demonstrate positive student growth, institutions must be able to accurately measure changes in these constructs over time. This accurate assessment of student growth can also aid in improving educational services. Programs that show evidence of positive student growth on a number of dimensions can be supported and expanded, whereas

programs that fail to nurture positive growth can be modified and improved. Thus, the accurate assessment of student growth is essential to meeting accountability demands while continually improving educational quality.

Despite the importance of accurate measurement of student growth over time, there are a number of practical issues that may reduce the accuracy of student growth estimates. For example, imagine you are an assessment coordinator for a mid-sized four-year university. University administrators want to ensure that student scientific reasoning skills are improving as a result of attending the university. To assess growth in scientific reasoning, you implement an assessment design where entering college students complete a scientific reasoning exam, and these same students are retested after completing the first three semesters of their coursework. If students' average scientific reasoning test scores increased between the pretest and the posttest, this increase would provide some evidence of the effectiveness of university science programming. As the exam is primarily designed to measure program effectiveness, you decide the exam will be low-stakes for students. That is, performance on the exam will have no personal consequences for the individual student (e.g., test score not factored into grades or associated with graduation). After collecting data for a number of years, you notice that a subset of students who completed the pretest and three semesters of coursework did not complete the posttest upon request. Unfortunately, you have little information to infer the exact reason *why* students are not completing the posttest. Although these students may have been sick the day of the posttest, another possibility is these students simply did not want to participate in the posttest, and hence "skipped" the test. No matter the cause for the missing posttest scores, you wish to address this missing data issue in a manner that does not bias

estimates of growth in scientific reasoning skills for students completing the first three semesters of university coursework.

The purpose of this study was to determine the best manner of handling missing data in an educational assessment context similar to the one described above. Prior to presenting the specific research questions for this study, I will review the issues surrounding missing data. As will be explained, the impact of missing data depends on the mechanism that resulted in the missingness. Unfortunately, this mechanism can only be empirically determined by knowing the values of the missing data. Although it is generally recommended to attempt to recover missing data by tracking and contacting missing participants (Glynn, Laird, & Rubin, 1993; Graham & Donaldson, 1993), researchers are often unable to do so due to budgetary or practical issues. Thus, the first goal of this study was to determine the exact mechanism underlying missingness due to posttest nonattendance by actually securing initially missing posttest scores. As Graham (2009) noted, "With a few well-placed studies of this sort, we would be in an excellent position to establish true bias from using [a variety of missing data] methods" (p. 571). Thus, after establishing the missing data mechanism, the second goal of this study was to determine the amount of bias introduced by various missing data handling techniques. More specifically, because the initially missing values were obtained via follow-up testing, the results using the complete dataset (i.e., including the initially missing scores) can be compared to the results obtained using various missing data handling techniques (i.e., excluding the initially missing scores). The manner in which these results can inform best practices for handling this type of missing data in future educational assessment will also be discussed.

Missingness in Educational Accountability Data

Missing data scenarios involving attrition over time are familiar to both higher education and K-12 assessment practitioners. For example, K-12 student participation rates for some National Assessment of Educational Progress (NAEP) assessments can be lower than 50 percent at later grade levels (i.e., 12th grade; Chromy, 2005). Moreover, the source of missingness is rarely investigated or reported in educational testing contexts (Amrein-Beardsley, 2008; Rubin, Stuart, & Zanutto, 2004). Rather, cases with missing data are often simply excluded from analysis.

Unfortunately, missing data can constitute a significant challenge to accurate inferences regarding student development and program effectiveness. Particularly, the common practice of excluding cases from analysis via *pairwise* or *listwise deletion* can introduce significant bias to parameter estimates and inflate standard errors. *Pairwise deletion* involves excluding cases from a specific analysis when data are missing for any variable involved in the given analysis. *Listwise deletion* involves excluding cases with any missing data from all analyses, regardless of whether the variables with missing data are involved in a particular analysis. As noted by Wilkinson and the Task Force on Statistical inference (1999), "[Listwise and pairwise deletion] are among the worst methods available for practical applications" (p. 598). In the example above, suppose only students with high scientific reasoning ability after three semesters complied with the request to complete the posttest. That is, students with low scientific reasoning ability avoided the posttest and account for the majority of the missingness at posttest. In this case, there is a *reason for* or *cause of* missingness: low scientific reasoning ability. Thus, missingness (attending vs. skipping the posttest) *depends on* the posttest scores (including

both the observed posttest scores and the posttest scores that would have been observed from the students who initially skipped the posttest). If the low-ability students' data were not included in the analysis, the growth estimate associated with scientific reasoning skills would likely be upwardly biased, primarily representing change in scores for the high-ability students. Additionally, standard errors would be inflated if the number of students skipping the posttest was large. In an alternative scenario, imagine the students missing at posttest were ill at the time of posttesting, and thus were no different with respect to scientific reasoning ability from the students for whom posttest data were observed. In this case, the missingness is *random* with respect to scientific reasoning scores, and the estimates of pre-post change may not be biased if the ill students were excluded from the analysis. However, standard errors would still be inflated if the number of ill students was large, due to the reduced sample size.

As highlighted in this hypothetical example, the reason, or mechanism, underlying the missingness can have a profound effect on the magnitude of the growth estimate. Thus, the mechanism underlying missingness impacts the appropriateness of different methods for analyzing the change in scores over time. If the missingness is truly random, traditional methods of handling missing data (e.g., listwise and pairwise deletion) will provide accurate estimates of change, although standard errors may be inflated. However, if the missingness is not random, estimates of change can be significantly biased if an inappropriate technique for handling the missingness (e.g., listwise and pairwise deletion) is employed. Thus, it is important to understand the different mechanisms that can result in missing data, as the missing data mechanism dictates the acceptable approach to handling the missingness.

Missing Data Mechanisms

Fortunately, researchers have investigated the conditions under which various parameter estimates may be biased due to missing data (e.g., Enders, 2010; Schafer & Graham, 2002). More specifically, Rubin (1976) developed a classification scheme for missing data mechanisms that is useful when considering how to appropriately account for missingness. Missing data mechanisms can be considered missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). Each missing data mechanism will be briefly reviewed below, followed by a description of how one should address each type of missingness during data analysis. A more detailed review of different data analytic techniques appropriate under these mechanisms is provided in Chapter 2. After outlining the missing data mechanisms below, the issue of missing posttest scores when assessing "value-added" for higher education accountability mandates will be further discussed. That is, plausible missing data mechanisms underlying missing posttest scores and the implications of those mechanisms will be presented.

What determines the missing data mechanism? Missing data mechanisms are not characteristics of the dataset. Rather, the mechanisms are assumptions associated with a specific analysis (Baraldi & Enders, 2012; Rubin, 1976). The mechanism underlying missingness is determined by the relationships between the missingness (*R*), the variable with missing data itself (*Y*), and other measured variables in the dataset (see Table 1). A missingness variable, *R*, can be computed by assigning a value of 0 to a case if *Y* is missing and a value of 1 if *Y* is observed. As is outlined below, the missing data mechanism is determined by whether *R* is related to the variable with missing data itself

(*Y*), other measured variables in the dataset (*X*), and whether *R* is related to *Y* conditional on the other measured variables in the dataset (*X*).

Missing completely at random (MCAR). The missing completely at random (MCAR) assumption is satisfied when missingness (*R*) on variable *Y* is unrelated to all measured variables in the dataset (*X*), as well as to *Y* itself (Enders, 2010). This mechanism is displayed in Figure 1a. For instance, suppose that scientific reasoning ability was measured for all incoming college students (i.e., pretest), but only a random sample of students were administered the exam three semesters later (i.e., posttest) due to cost concerns (e.g., pencils, paper, proctors). In this case, *R* would be completely random, by design, and would therefore be unrelated to both *Y* and all other variables in the dataset. This design is known as a *planned missingness design*, and is one of the most common missing data scenarios that result in the MCAR assumption being met. However, it is also possible to meet the MCAR assumption when missingness is unplanned. For example, if some students miss posttest due to illness, it is likely that missingness (*R*) would be unrelated to any measured variables in the dataset, and also unrelated to *Y*, resulting in the missingness meeting the MCAR assumption.

Missing at random (MAR). The missing at random (MAR) assumption is satisfied when missingness (*R*) on variable *Y* is unrelated to *Y* itself *after controlling for the other measured variables included in the analysis* (Heitjan & Basu, 1996). That is, *R* may be bivariately related to *Y*, but this relationship is *spurious* and does not remain significant after controlling for other variables included in the analysis. Thus, the MAR assumption is more relaxed than the MCAR assumption. This mechanism is displayed in Figure 1b.

Unlike MCAR, the MAR mechanism indicates there is a variable or set of variables that explains missingness. For example, suppose that students who scored below a certain threshold on the scientific reasoning pretest were expelled from the university. The remaining students then completed the scientific reasoning posttest. If the expulsions were the only reason for missing posttest scores, missingness at posttest (*R*) could be completely predicted from (i.e., explained by) pretest scores (*X*). Although missingness (*R*) is likely related to the hypothetical complete set of posttest scores (*Y*) (i.e., including posttest scores that were observed and those that would have been obtained, but were instead missing), this relationship is completely explained by student pretest scores. Thus, *after controlling for* pretest scientific reasoning scores (*X*), missingness (*R*) would be unrelated to posttest scores (*Y*), thus meeting the MAR assumption. Note that measured variables in the researcher's dataset do not need to completely predict missingness for the mechanism to be considered MAR. Rather, measured variables only need to predict the missingness *that is related to the variable with missing values* (*Y*). For example, suppose that, in addition to expelling students with low pretest scores, a number of students also missed posttest due to reasons unrelated to their scientific reasoning scores (e.g., some students were sick). In this case, pretest scores (*X*) would not perfectly correlate with missingness variable *R*. However, pretest scores would account for the portion of *R* that is associated with posttest scores (*Y*), and thus the posttest data should be considered MAR, as *R* is unrelated to *Y* after controlling for pretest scores (Baraldi & Enders, 2012).

Missing not at random (MNAR). The MNAR mechanism occurs when data are missing in a manner that is related to the variable with missing data itself after controlling

for other variables in the dataset. For example, suppose that pretest scores were not collected or not included in the data analysis in the previous expulsion scenario. Referring to Figure 1b, the pretest score (*X*) would not be included in the figure. In this case, missingness (*R*) and posttest score (*Y*) would be significantly related (i.e., the dashed curve arrow representing the correlation between *R* and *Y* in Figure 1b would no longer be approximately zero, but would be some non-negligible value).

MNAR data can also result if the other measured variables included in the analysis (*X*) do not fully explain the relationship between missingness (*R*) and posttest score (*Y*), as is displayed in Figure 1c. For example, in addition to missingness being due to low pretest scores (*X*), suppose that some students fail to attend the scientific reasoning posttest due to low academic self-efficacy. These students would likely score lower on the scientific reasoning posttest, so missingness (*R*) is related to posttest scores (*Y*). Pretest score (*X*) does not completely explain the relationship between missingness (*R*) and posttest score (*Y*). That is, *R* remains related to *Y*, even after controlling for other measured variables in the dataset, thus reflecting a MNAR mechanism. In Figure 1c, the curved arrow connecting *R* and *Y* represents a non-negligible relationship between missingness (*R*) and posttest scores (*Y*), even after controlling for pretest (*X*). In this example, the curved arrow represents the relationship between missingness (*R*) and posttest scores (*Y*) due to their shared relationship with self-efficacy. If self-efficacy was measured *and included in the analysis*, one could satisfy the MAR assumption by accounting for the relationship between missingness and posttest scores; that is, missingness would no longer be related to posttest scores after partialling out the variance due to self-efficacy. Thus, this example highlights that missing data mechanisms are not a

characteristic of the dataset, but rather are assumptions associated with the specific analysis being conducted (Baraldi & Enders, 2012).

Determining the missing data mechanism. Further complicating researchers' and assessment practitioners' attempt to account for missing data is it is usually impossible to determine the exact mechanism underlying missingness (Table 1). Recall the missing data mechanism is determined by whether missingness (*R*) is related to other measured variables in the dataset, *and* whether *R* is related to the variable with missing values (*Y*), conditional on other measured variables (*X*s) included in the analysis. The relationship between *R* and all other measured variables can be directly estimated and evaluated for statistical significance. If *R* relates significantly to any measured variable (*X*), the MCAR assumption is falsified, and the missingness mechanism must be considered either MAR (if *R* is unrelated to *Y* after controlling for *X* variables) or MNAR (if *R* remains related to *Y* after controlling for *X* variables). By contrast, if *R* does not significantly relate to any measured variable, then no measured variable can moderate the relationship between *R* and *Y*. Thus, the missingness mechanism data must be considered either be MCAR (if *R* is unrelated to *Y*) or MNAR (if *R* is related to *Y*).

However, *Y* will be missing for all cases where $R = 0$. Consequently, the relationship between *R* and *Y* cannot be empirically estimated, as this would require the missing scores. Thus, even if *R* is found to be unrelated to other measured variables, there is no way to empirically determine if *R* is related to *Y*. Consequently, the MCAR mechanism is empirically indistinguishable from the MNAR mechanism. Similarly, if *R* is found to relate to other measured variables, there is no way to determine if *R* is related to *Y* after controlling for the other measured variables in the dataset. In this case, the

MAR mechanism is empirically indistinguishable from the MNAR mechanism. Thus, unless missingness is carefully planned, the MNAR mechanism is always a possibility that cannot be empirically falsified.

Although the exact missing data mechanism can rarely be empirically determined, researchers and practitioners may be able to infer the mechanism. For example, researchers and practitioners may assume MCAR if a planned missingness design was properly implemented and all missingness was a result of that design. For unplanned missingness, researchers might locate and interview a few respondents that had missing data and determine their reasons for missingness (Enders & Gottschall, 2011). If the reasons seem to be related to the missing variable values themselves, and unrelated to other measured variables in the dataset, a MNAR mechanism is likely to underlie the data. If the reasons for missingness seem to be unrelated to any variables of interest (e.g., illness), then a MCAR mechanism may be plausible.

General Recommendations for Handling Missing Data

There are two general approaches to addressing missing data issues. The first approach is to avoid the problem of missingness entirely by observing data that would have otherwise been missing. This approach can be done preventatively by adopting a research design that limits attrition. Examples of attrition prevention strategies include decreasing participant burden, increasing participant incentives, increasing contact with participants, or changing the timing of measurement occasions in longitudinal designs (McKnight, McKnight, Sidani, & Figueredo, 2007). Additionally, multiple researchers recommend maintaining accurate and complete participant contact information to track and contact participants who have not provided data (Lavori, 1992; McKnight et al.,

2007). Alternative arrangements can be made to accommodate participant schedules and recover data that would have otherwise been missing (Glynn et al., 1993; Graham & Donaldson, 1993). In an educational testing environment, this strategy may include having multiple testing sessions to allow students to attend different testing times. In the current study, the initially missing posttest scores were recovered via a makeup testing session. Thus, complete data were obtained and the exact missing data mechanism can be empirically determined.

Unfortunately, the prevention or recovery of missing data may not always be possible. Thus, the second approach to addressing missing data is to incorporate the missingness into data analysis. Most missing data researchers recommend an *inclusive data analysis strategy* to deal appropriately with missing data, regardless of the mechanism of missingness (Collins, Schafer, & Kam, 2001; Enders, 2010; Rubin, 1996; Schafer, 1997; Schafer & Graham, 2002). This strategy involves measuring a number of variables that are hypothesized to relate to either missingness (*R*) or the variable for which missingness is present (*Y*). These variables (*X*s) are then included as *auxiliary variables* in the analysis of the data using multiple imputation (MI) or full information maximum likelihood (FIML) estimation. Referring to Figure 1b, an auxiliary variable (*X*) was incorporated into the analysis of *Y* to address missingness. Although the specifics of MI and FIML are different, both techniques utilize the relationships between *R*, *Y*, and the auxiliary variables (*X*s) to better estimate parameters involving *Y*. The auxiliary variables may not be of substantive interest to the researcher, but are rather used to aid in estimation of parameters associated with the variable with missingness (i.e., used to aid in the estimation of parameters that are of substantive interest).

Utilizing an inclusive data analysis strategy can allow data that should be considered MNAR to meet the MAR assumption (Collins et al., 2001; Savalei & Bentler, 2009). Referring to Figure 1b, incorporating auxiliary variables (*X*) that are related to missingness (*R*) and the variable with missing values (*Y*) increases the likelihood that missingness and the variable with missing values will not be significantly related after controlling for the auxiliary variables (*X*). Thus, a MNAR mechanism can be transformed into an MAR mechanism with the inclusion of auxiliary variables. In this manner, adopting an inclusive data analysis strategy reduces the likelihood that a MNAR mechanism underlies the data and increases the likelihood that the missingness will meet the MAR assumption.

The utilization of an inclusive data analysis strategy, combined with MI or FIML estimation, appears to be the best analysis alternative under the majority of missing data scenarios. Under MAR conditions, the inclusive data analysis strategy produces more accurate parameter estimates than excluding auxiliary variables (Collins et al., 2001). Further, the strategy reduces standard errors under both MAR and MCAR conditions. Unfortunately, the inclusive data analysis strategy still results in biased parameter estimates under a MNAR mechanism. However, MNAR-based methods often require strong assumptions regarding the missingness. If these assumptions are not met, the results of the MNAR-based analyses can lead to worse estimates than the MAR-based inclusive data analysis approach (Demirtas & Schafer, 2003). Thus, researchers have argued that MNAR-based strategies should not be routinely used (Enders, 2010; Schafer & Graham, 2002). Given 1) there is currently no practical method to account for MNAR data statistically in most missing data situations, 2) an inclusive data analysis strategy

limits parameter bias and standard errors under MAR and MCAR conditions, and 3) in the typical research or testing setting one never knows the exact missing data mechanism, this inclusive data analysis approach to handling missing data is usually recommended if missing values cannot be recovered.

Missing Data Handling Practices in Educational Assessment

Given an inclusive data analysis strategy appears to be the best way to handle missingness in the majority of missing data scenarios, one would hope this strategy is commonly used when examining student development for institutional accountability purposes. Unfortunately, institutions often use listwise or pairwise deletion when faced with missing data. For example, many value-added statistical models in K-12 accountability testing are applied to only complete cases, thus listwise deleting any cases with missing data (Amrein-Beardsley, 2008; Rubin et al., 2004). "Given the large proportion of missing data in many achievement databases and known differences between students with complete and incomplete test data, it is possible that estimates may be highly sensitive to this (or other) assumptions about missing data" (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004, p. 97). Given that students with missing data on many K-12 assessments tend to be low-performing (Amrein-Beardsley, 2008), a MCAR mechanism, which listwise deletion assumes, is extremely unlikely.

Higher Education Accountability Data Examined in the Current Study

An MCAR mechanism was similarly unlikely to underlie the missingness in the higher education accountability data being examined in the current study. At this mid-sized mid-Atlantic public university, students are measured at two time points to assess the effectiveness of general education and student affairs programming. All students are

tested initially as incoming first-year students and again after they have accumulated between 45 and 70 credit hours. All university classes are cancelled for these "Assessment Days." Students are randomly assigned to rooms based on their university-assigned student identification numbers and receive different testing configurations based on room assignment. These testing configurations include both cognitive (i.e., knowledge-based) and noncognitive (i.e., attitude-based) assessments. In this manner, the assessments utilize a planned missingness design; not all students complete every instrument, but the random assignment of students to different testing configurations ensures that the missingness due to not receiving an instrument is completely random. Although the students at the second testing session (i.e., posttest) completed either three or five semesters of coursework at the university, only students completing three semesters of coursework are of interest in computing student growth estimates. That is, university administrators are chiefly interested in the change in cognitive and noncognitive constructs experienced by students completing between 45 and 70 credit hours within the first three semesters of university attendance. Thus, test configurations are matched between the first-year student assessment sessions held during a given Fall semester (i.e., pretest) and the assessment sessions held during Spring three semesters later (i.e., posttest). The university attempts to assign students to the same testing room for their second testing session, so that pre-post change can be examined on the constructs of interest.

Although students are required to attend their assigned Assessment Day testing sessions, there are no personal consequences tied to individual performance on the tests. That is, the testing is low stakes for students. Every year, there are a number of students

who fail to attend their assigned assessment session. Given that the first (i.e., pretest) Assessment Day is integrated into the university orientation program, nonattendance is typically minimal at pretest and much more common at the second (i.e., posttest) Assessment Day. To compel nonattending students to participate, the university places registration holds on the students' accounts. This academic hold prevents students from registering for classes until they attend a makeup assessment session. These sessions are held on a Friday evening or Saturday morning. Via these makeup sessions, the university is able to eventually test every student, aligning with the recommendations to avoid missing data issues by recovering data from students who initially did not provide data (McKnight et al., 2007). However, the university currently does not include the makeup data when computing value-added estimates. Specifically, the value-added estimates are computed using only those students who provided scores at both pretest and posttest Assessment Day testing sessions. Thus, although the university is subsequently gathering the "missing data" via the makeup testing sessions, the data is not included in analyses, potentially resulting in biased estimates and inflated standard errors.

Fortunately, this data collection scheme (i.e., posttest data collected from students who were initially missing at posttest) allows for the investigation of missing data issues in accountability testing. In addition to uncovering the missing data mechanism, the parameter estimates and standard errors obtained from the complete dataset (i.e., including posttest data obtained from makeup sessions) can be compared to parameter estimates and standard errors obtained when treating makeup data as missing and utilizing different missing data analysis techniques. The different datasets available are displayed in Figure 2. Currently, the scores of students with makeup posttest data are

listwise deleted from analyses (Dataset 1). Under a MCAR mechanism, excluding the makeup students from analysis should result in unbiased average student growth estimates. In this case, the growth estimates obtained excluding makeup students should be comparable to the growth estimates obtained from the complete dataset that includes makeup students (Dataset 2). However, even under a MCAR mechanism, standard errors may be inflated when excluding the makeup students due to the decreased sample size (note how analyses of Dataset 1 are based on four students, whereas analyses of Dataset 2 are based on six students). Additionally, under a MAR mechanism, excluding the makeup students from analyses would produce biased growth estimates. Instead of listwise deleting students who attended posttest makeup testing, an alternative method of handling this "missing" posttest data would be to utilize MI or FIML techniques (Dataset 3; analyses would be based on all six students even though Students 5 and 6 don't have posttest scores). Under a MAR mechanism, adopting an inclusive analysis strategy combined with MI or FIML techniques should result in growth estimates that are closer to those obtained from the complete data (Dataset 2) than simply deleting students with missing posttest values (Dataset 1). Finally, under a MNAR mechanism, both listwise deletion (Dataset 1) and the inclusive analysis strategy (Dataset 3) would result in biased estimates of student growth, but the inclusive analysis strategy should result in decreased bias and standard errors relative to listwise deletion.

Possible missing data mechanisms underlying Assessment Day nonattendance. It is important to understand, to the extent possible, the reasons *why* students do not attend the second Assessment Day (i.e., posttest), and thus must attend a makeup session. That is, understanding the correlates of non-attendance (*R*) can help

identify whether the missing posttest data (*Y*) should be considered MCAR, MAR, or MNAR. Understanding the missing data mechanism would be valuable in situations when the posttest data cannot be collected via makeup testing, *or if collected but not included in data analyses (as is current practice)*. In short, establishing the missing data mechanism underlying the initially missing data (i.e., makeup data) can help inform the best way to handle the data.

Previous research indicates that a MCAR mechanism is implausible. Makeup examinees have been found to be qualitatively different from examinees who attend Assessment Day. Students who skip assessment day are more likely to be male and less motivated to perform well on assessments (Swerdzewski, Harmes, & Finney, 2009). Importantly, students who skip Assessment Day score significantly lower on cognitive tests (Swerdzewski et al., 2009). That is, there is evidence that missingness (*R*) is related to posttest scores on cognitive tests (*Y*), ruling out the MCAR mechanism. The extent to which the mechanism is considered MAR or MNAR would depend on the auxiliary variables (*X*) measured in a given year (e.g., gender), whether these variables are included in the data analysis, and the extent to which these variables moderate the relationship between missingness (*R*) and posttest scores (*Y*).

Given the makeup posttest data are unlikely to meet the MCAR assumption, the current method of analyzing accountability data at this university is problematic. That is, the listwise-deletion used by university assessment specialists to handle student makeup data is only appropriate under MCAR conditions. Thus, this method may be introducing bias into student growth estimates. However, both the specific missing data mechanism

(i.e., MAR or MNAR) and the degree of bias introduced by excluding makeup student data from analyses are unclear.

Purpose of the Current Study

The current study aimed to uncover 1) the missing data mechanism (i.e., MCAR, MAR, or MNAR) associated with low-stakes testing attrition and 2) the impact of employing different missing data techniques on value-added estimates and their associated significance tests. The assessment data used for the current study were unique in that data were recovered from students who were initially missing at posttest. Given that the "missing" values are known, the missing data mechanism (i.e., MCAR, MAR, or MNAR) can be empirically identified. That is, "missingness" (i.e., *R*, whether the posttest score was collected during Assessment Day or during a makeup session) can be correlated with the values of the "missing" posttest data (*Y*), both before and after controlling for the other measured variables in the dataset (i.e., auxiliary variables). If this *R*-*Y* relationship is found to be significant *without* auxiliary variables (indicating an MNAR mechanism), but non-significant *when including* auxiliary variables (indicating a MAR mechanism), this would indicate that the MAR assumption would only be met *when auxiliary variables are included in analyses*. Interestingly, if the data are found to be MNAR, the *extent* to which the data can be considered MNAR can also be examined. That is, missingness (*R*) may be statistically significantly related to the missing data values (*Y*) after controlling for auxiliary variables, but only weakly. In this case, the MNAR mechanism would be expected to bias results less drastically when utilizing auxiliary variables in a MAR-based analysis (e.g., MI or FIML) than if missingness (*R*)

was strongly related to the missing data values (*Y*) after controlling for auxiliary variables.

After identifying the missing data mechanism, the value-added estimates obtained using the complete data (Dataset 2) were compared to value-added estimates obtained if students with missing data are excluded (Dataset 1) and value-added estimates if posttest data from make-up examinees are treated as missing (Dataset 3) using different missing data handling techniques. The differences between these results can inform best practices for assessment practitioners encountering this form of missingness in the future. For instance, if the parameter estimates and standard errors obtained by excluding students with makeup posttest data (Dataset 1), or by utilizing any of the modern missing data handling techniques (analyzing Dataset 3 using MI or FIML) are comparable to those obtained by analyzing the complete data (Dataset 2), this may indicate that the current practice of excluding students with makeup posttest data is acceptable and does not result in significant bias or loss of power. If utilizing the missing data handling techniques (analyzing Dataset 3 using MI or FIML) result in parameter estimates and standard errors that are comparable to those obtained by analyzing the complete data (Dataset 2), but excluding the makeup students (Dataset 1) results in bias or loss of power, this would indicate that future assessments should utilize MI or FIML. Finally, if excluding students with makeup posttest data (Dataset 1) and utilizing modern missing data handling techniques to account for posttest missingness (Dataset 3) both result in substantial bias or loss of power compared to analyzing the complete data (Dataset 2), this would indicate utilizing the makeup assessment data is essential to obtaining accurate assessment results. Thus, the results of this study can provide valuable guidance for assessment practice.

This research design has a number of advantages over previous simulation and applied missing data analysis studies. Unlike simulated data, the data used in this study were collected in a real missing data scenario. Unlike typical applied missing data analysis studies, the values of the "missing" data are known (due to recovering the initially missing data via a makeup session). Thus, the true relationship between missingness (*R*) and the variable with initially missing values (*Y*) could be estimated, and the missing data mechanism in an operational testing program could be empirically determined. After establishing the missing data mechanism, the results obtained utilizing various missing data handling techniques could be compared to the results obtained using the complete dataset. Through this comparison, the extent of bias introduced by missingness could be empirically assessed in a *real* data situation, which is valuable to the study of attrition in low-stakes educational testing settings. In this manner, addressing the research questions outlined below facilitates a better understanding of the causes and effects of missingness on pre-post change estimates obtained from educational accountability data, and informs best practices on the handling of such missingness. The specific implications associated with each research question are presented below.

Research Question 1: Examining posttest response validity. To what extent can the posttest scores provided by students in the makeup testing sessions be considered valid? Before investigating the mechanism underlying posttest nonattendance, it was important to determine the extent to which the students attending the makeup testing session at posttest provided valid responses. That is, students providing data at makeup testing report putting forth less test-taking effort than students attending the assigned Assessment Day session (Swerdzewski et al., 2009), and could thus be providing invalid

responses at posttest by responding randomly. In this case, student growth estimates obtained by including the makeup students in the analysis could be considered *biased*, as the estimates would not be reflective of the true growth in student knowledge, skills, or abilities. If makeup students are responding randomly at posttest, the prediction of posttest scores from pretest scores should be different for makeup students when compared to students attending Assessment Day at posttest. That is, when regressing posttest scores on pretest scores, the intercept, slope, or unexplained posttest variance would differ between Assessment Day and makeup students if the students attending a makeup session did not provide valid posttest responses. More specifically, random responding by makeup students may reduce the pretest-posttest slope or increase the unexplained variance in posttest scores. Additionally, less posttest effort by makeup students may also reduce the average posttest score, resulting in a reduced intercept for makeup students when compared to Assessment Day students. These possibilities were investigated to ensure that parameters obtained utilizing the complete (i.e., including posttest makeup) dataset were accurate reflections of overall student growth, and were not biased by the inclusion of makeup student data.

Research question 2: Examining the missing data mechanism. What missing data mechanism underlies the initially missing posttest data (i.e., posttest makeup data)? "Missingness" in this study refers to whether a student attended their assigned assessment session at posttest, or if they were instead compelled to attend a makeup assessment session. This dichotomous "missingness" variable could be: 1) unrelated to other measured (i.e., auxiliary) variables, as well as unrelated to posttest scores (i.e., a MCAR mechanism); 2) related to other measured variables, but unrelated to posttest scores after

controlling for the other measured variables (i.e., a MAR mechanism); or 3) related to posttest scores after controlling for all other measured variables (i.e., a MNAR mechanism). Given that "missing" posttest scores were obtained from students completing a makeup assessment at posttest, the missing data mechanism could be empirically determined, which would be impossible in most applied missing data scenarios.

As mentioned previously, research has found that Assessment Day non-attendance is related to a number of student attributes (Swerdzewski et al., 2009; Zilberberg, 2013). Thus, it appears that assuming a MCAR mechanism is unjustified. However, this study further investigated whether the makeup data, if treated as missing, should be considered MAR or MNAR. That is, MAR and MNAR mechanisms are distinguished by whether "missingness" (i.e., whether a student attended Assessment Day or a makeup session at posttest) is related to posttest outcome scores (e.g., scientific reasoning), *after controlling for other measured variables included in the analysis*.

Determining the precise mechanism underlying missingness has implications for higher education accountability testing practice. That is, the current method of listwise deleting the scores of makeup students would only be appropriate if a MCAR mechanism is found to underlie the missingness. However, if a MAR mechanism were identified, the university should abandon listwise deletion and utilize MI or FIML with auxiliary variables to more accurately estimate average student growth. Additionally, the ability to investigate the actual missing data mechanism allows for the identification of salient auxiliary variables that should be used in the estimation of student growth estimates in the future. If a MNAR mechanism were found to underlie the data, then the makeup data

should be *included* when computing average student growth estimates. That is, the other variables measured as a part of university assessment cannot account for the effects of excluding makeup students' data from value-added estimates. Thus, these students' posttest scores must not only be gathered but also included in data analysis to accurately measure student growth. In addition to informing practice at this particular university, other testing programs utilizing a low-stakes, pre-post assessment design would likely have missingness of the same nature (e.g., NAEP data, Chromy, 2005). Thus, the results of this study may provide guidance regarding how missing data should be handled at other institutions with similar missing data issues.

Research question 3: Comparing missing data handling techniques. How do the estimates of growth differ across the methods of handling the missing data, and how do these results compare to those obtained from combining the Assessment Day and makeup posttest data to create the complete dataset? That is, posttest data were obtained from students during makeup testing sessions that would have been missing if those makeup sessions were not conducted. Thus, the results obtained from the complete dataset (including makeup student posttest data; Dataset 2 in Figure 2) can be compared to results that would be obtained if the makeup student posttest data are treated as missing (Datasets 1 and 3 in Figure 2) using different missing data handling techniques.

To answer this question, multiple missing data techniques were utilized, and the results were compared to those obtained from the complete dataset. Mean pre-post growth estimates, in addition to mean posttest scores, the variance of the posttest scores, and the covariance of the pretest and posttest scores, were obtained from eight methods:

1) Utilizing the complete dataset, which includes makeup posttest data (i.e., including those who were initially missing by recovering their scores via "makeup" testing; Dataset 2 in Figure 2)
2) Utilizing listwise deletion, excluding examinees that attended makeup testing sessions at post-test (Dataset 1 in Figure 2)
3) Treating makeup posttest data as missing and utilizing multiple imputation (MI) without auxiliary variables (Dataset 3 in Figure 2)
4) Treating makeup posttest data as missing and utilizing MI with university database and pretest auxiliary variables (Dataset 3 in Figure 2)
5) Treating makeup posttest data as missing and utilizing MI with all auxiliary variables (Dataset 3 in Figure 2)
6) Treating makeup posttest data as missing and utilizing full information maximum likelihood (FIML) without auxiliary variables (Dataset 3 in Figure 2)
7) Treating makeup posttest data as missing and utilizing FIML with university database and pretest auxiliary variables (Dataset 3 in Figure 2)
8) Treating makeup posttest data as missing and utilizing FIML with all auxiliary variables (Dataset 3 in Figure 2)

Note that Method 1 is the most desirable assessment design, as complete data is gathered and used in the estimation of pre-post growth. Method 2 is currently being used by the university, but is generally not recommended by missing data experts (Enders, 2010; Wilkinson & Task Force, 1999). Methods 3 - 8 exclude posttest makeup data, but pretest data are included and aid in the estimation of growth estimates. Importantly, MI and FIML analyses were conducted multiple times with different sets of auxiliary variables.

As mentioned previously, recommended auxiliary variables are variables that are associated with missingness, values of the missing variable itself, or both (Enders, 2010). Thus, any variable associated with Assessment Day posttest attendance or posttest scores could be considered a potential auxiliary variable.

The extent to which the inclusion of auxiliary variables reduces bias and standard errors *depends on the nature of the relationships between the auxiliary variables, missingness, and posttest scores*. Table 2 summarizes the effect of excluding auxiliary variables under particular conditions, as determined by Collins and colleagues (2001). In brief, including an auxiliary variable (*X*) that is unrelated to posttest scores (*Y*) should not affect parameters or standard errors associated with posttest mean, posttest variance, pretest-posttest covariance, and pre-post mean difference. Including a variable (*X*) that is related to posttest scores (*Y*) but unrelated to missingness (*R*) should result in unaffected parameter estimates, but reduced standard errors. Including an auxiliary variable that is related to posttest scores and linearly related to missingness should result in reduced bias in parameter estimates and reduced standard errors. Finally, including an auxiliary variable that is related to posttest scores and nonlinearly related to missingness should result in reduced bias in posttest variance and pretest-posttest covariance estimates, as well as reduced standard errors, but unaffected posttest mean and pre-post mean difference estimates. These effects should be more pronounced for auxiliary variables that are more strongly related to posttest scores. MI and FIML analyses without auxiliary variables still included pretest scores in the estimation of pre-post growth, and thus should produce more accurate growth estimates than listwise deletion under a MAR

mechanism. The effect of including different types of auxiliary variables on parameter estimates and standard errors is discussed in more detail in Chapter 2.

Why compare the results produced when employing different sets of auxiliary variables? This comparison should indicate the utility of including different sets of auxiliary variables to obtain more accurate growth estimates. That is, assessment practitioners may not have access to a wealth of student information to utilize as auxiliary variables. In some cases, the only data available to assessment practitioners may be the students' pretest and posttest scores. Additionally, an assessment practitioner choosing to omit makeup testing in favor of utilizing missing data handling techniques *would not have access to posttest auxiliary variables.* That is, the posttest auxiliary variables are collected *during the posttest*, and thus would not be available for makeup examinees if makeup testing were not conducted. Thus, it was important to compare the performance of the MAR-based missing data procedures (MI and FIML) without auxiliary variables, with only university database and pretest auxiliary variables, and with all auxiliary variables, as this comparison may highlight the necessity of gathering particular auxiliary variables. Previous research indicates that results are generally improved by the inclusion of auxiliary variables (Collins et al., 2001). Thus, compared to MI and FIML procedures excluding auxiliary variables, including auxiliary variables should produce growth estimates closer to those obtained using the complete dataset. Additionally, directly comparing these methods with and without auxiliary variables should give an indication of the *degree* to which results are improved by including certain sets of auxiliary variables. If including auxiliary variables provides pre-post growth estimates that are much closer to those obtained using the complete dataset, assessment practitioners should

spend additional time and resources collecting that auxiliary variable data as a part of their assessment design. However, a negligible difference in pre-post growth estimates with and without auxiliary variables would indicate that auxiliary variable data collection may not be worth the additional cost.

Some may question the utility of examining *both* MI and FIML results, given both are designed for MAR data and provide similar results. As will be explained in Chapter 2, the methods by which MI and FIML estimate parameters are mathematically different. For instance, auxiliary variables are included in the MI procedure via an imputation model that is separate from the analysis model, whereas auxiliary variables must be integrated into the analysis model in the FIML procedure. Including a large number of auxiliary variables in FIML analyses may cause estimation difficulties (Savalei & Bentler, 2009). Thus, it is important to compare MI and FIML results to uncover potential difficulties that may be associated with one technique, but not the other. Additionally, MI provides multiple datasets with imputed posttest scores. If the parameter estimates (posttest mean, posttest variance, pretest-posttest covariance, and pre-post mean difference) obtained utilizing MI differ widely from those obtained utilizing the complete dataset, the individual imputation values can be examined to determine the extent to which they differ from the actual values in the complete dataset. This examination may help identify outliers, or individual students with actual posttest scores that are substantially different from their imputed posttest scores. For example, some makeup session students may have aberrantly low posttest scores due to lower test-taking motivation (Swerdzewski et al., 2009). Students with actual scores that are substantially different from their imputed scores can also be closely examined to identify additional

auxiliary variables. For instance, suppose that a disproportionate number of international students had actual posttest scores that were substantially different from their imputed posttest scores. In this case, international student status should be included as an auxiliary variable. Or, in the example above, if low test-taking motivation is associated with disparate actual and imputed posttest scores, then test-taking motivation should be added as an auxiliary variable. In short, closely examining the imputed MI posttest scores could provide a wealth of information beyond examining FIML results.

The implications of differences in the results obtained via these data analysis methods inform assessment practice. If results are similar across the different methods, any of the methods can be used to obtain accurate growth estimates. However, if some methods of handling missingness outperform others by yielding growth estimates closer to those obtained from the complete dataset, then those methods should be used at this university and other institutions with similar missing data issues. Finally, if no method for handling missingness yielded growth estimates comparable to those obtained using the complete data, it may be necessary to obtain makeup data from students and to use this makeup data in pre-post growth analyses. As emphasized by Graham (2009), comparing the results of various missing data analysis techniques to those obtained utilizing the complete dataset in a real missing data scenario can inform the study of attrition in general, by examining the effects of *real* (not simulated) attrition on growth estimates.

Research question 4: Percentage of missingness. How are the previous results affected at varying proportions of missingness? At the university where this study was conducted, Assessment Day nonattendance is not extremely common. Currently, less

than 10% of students fail to attend their regularly scheduled Assessment Day testing session at posttest, and are forced to attend a makeup testing session. However, the percentage of missingness must be considered together with the missing data mechanism. That is, relatively small percentages of MNAR missingness could bias parameter estimates, whereas large percentages of MCAR missingness may have little effect on parameter estimates (but would still result in inflated standard errors) (Enders, 2010). However, the relatively small percentage of missingness present in this study may cause the effects of missingness on parameter estimates to be subtle.

Other institutions may have a greater proportion of missing data. For instance, given the same missing data mechanism, an institution with 50% student non-attendance at posttest would likely have growth estimates that are more biased than a university with 10% missingness. High missing data rates can be common in some testing programs, such as data collected for NAEP assessments (Chromy, 2005). In these cases, the handling of this missingness can have a profound effect on the results obtained from analysis of assessment data. Thus, the answers to the previous research questions were investigated at varying proportions of missingness.

Research question 5: Noncognitive vs. cognitive. Do the results of the previous research questions differ depending on whether growth is being estimated for noncognitive (e.g., developmental) or cognitive (e.g., scientific reasoning) constructs? Previous research indicates that students attending makeup testing are less likely to put forth effort on cognitively-taxing tests than on noncognitive developmental surveys, resulting in diminished performance on cognitive tests (Swerdzewski et al., 2009). Thus, the association between Assessment Day attendance (R) and posttest scores (Y) could be

stronger for cognitive tests (with lower posttest scores for makeup students) than for noncognitive measures. A stronger relationship between missingness (*R*) and the missing values themselves (*Y*) would indicate that a MNAR mechanism is more likely for cognitive tests than for noncognitive measures. Thus, it was important to investigate differences in results between noncognitive and cognitive measures.

A difference in results obtained when examining noncognitive vs. cognitive pre-post growth would indicate that different methods of handling missingness may need to be utilized depending on the construct being studied. For example, suppose that noncognitive makeup data met the MAR assumption, whereas cognitive makeup data did not and was thus considered MNAR. In this case, assessment practitioners could utilize MI or FIML with appropriate auxiliary variables when examining noncognitive constructs, but would need to obtain the complete data for cognitive constructs. Thus, examining these differences is important to inform best assessment practice.

CHAPTER TWO

Literature Review

Missing Data Techniques

The appropriateness of various techniques to account for missing data depends on the mechanism underlying the data. Rubin (1976) was the first researcher to develop a classification scheme to better understand missing data mechanisms. In addition to the variable with missing data, denoted *Y*, Rubin (1976) also defined a missingness variable, *R*. *R* is a binary variable that takes a value of 1 for cases where variable *Y* is observed, and takes a value of 0 for cases where variable *Y* is missing. Rubin (1976) defined data as missing-at-random (MAR) if missingness variable *R* is unrelated to *Y*, conditional on other observed data. However, if *R* is related to *Y* after controlling for other observed data, the data are considered to be missing-not-at-random (MNAR). The relationship between *R* and *Y* cannot be empirically estimated with applied data, given that *Y* is missing for all cases where $R = 0$. Rubin (1976) also defined data as observed-at-random (OAR) if missingness variable *R* is unrelated to the other observed data (i.e., variables other than *Y*). Data that are both OAR and MAR are considered missing-completely-at-random (MCAR; Heitjan & Basu, 1996).

Methods for Dealing with Missing Data

The methods outlined below and general recommendations regarding these methods are summarized in Table 3.

Deletion methods. Listwise and pairwise deletion are extremely common methods for handling missing data (Peugh & Enders, 2004). Despite their ubiquity, these methods are considered some of the worst for dealing with missing data (Little & Rubin,

2002; Wilkinson & Task Force on Statistical Inference, 1999). These methods assume that data meet the MCAR assumption. Deletion-based methods can significantly bias parameter estimates when the MCAR assumption is not upheld (Brown, 1994; Enders, 2001; Enders & Bandalos, 2001). Even under MCAR conditions, data deletion is wasteful and results in inflated standard errors. Given that methods that yield more parameter estimates and reduced standard errors are now available, deletion-based methods are not generally recommended (Enders, 2010).

Listwise Deletion. Listwise deletion involves deleting cases with any missing data on any variable. There are a number of benefits to listwise deletion. First, listwise deletion results in very low non-convergence rates (Enders, 2001; Enders & Bandalos, 2001). That is, it may be difficult for many software packages to estimate complex models involving many different variables with varying degrees of missingness. Analyzing only complete cases can ease the computational burden involved in model estimation. Second, listwise deletion greatly increases the practical ease of analysis, as no further treatment of missing data needs to be applied after removing cases with missingness. Third, listwise deletion has been found to yield unbiased estimates of association between two variables if the data meet the MCAR assumption.

Despite the benefits, there are two major problems associated with listwise deletion. First, removing cases results in decreased power and increased standard error estimates. The researcher is essentially "throwing away" information by needlessly deleting cases. This decreased power becomes more of an issue as the percent of missing data increases. Second, parameter estimates are biased when the assumption of MCAR is not met. For example, consider the situation where students complete a scientific

reasoning exam as entering freshmen, then again after three semesters of coursework. Suppose students with low pretest scores perform poorly at the university, and drop-out as a result. If the scores of these individuals were listwise deleted, mean posttest estimates of scientific reasoning ability would likely be too high, given that all students with low pretest scores were excluded at posttest. Given these substantial drawbacks, listwise deletion should not be used in the majority of missing data situations.

Pairwise deletion. Pairwise deletion involves excluding cases from analysis that are missing on the variables being analyzed. For instance, consider examining the relationship between three variables: *X, Y,* and *Z*. When estimating the correlation between variables *X* and *Y*, the researcher would exclude cases that had missing values for *X* or *Y*, regardless of whether data were missing on variable *Z*. Similarly, when examining the relationship between *X* and *Z*, the researcher would exclude cases that had missing values for *X* or *Z*, regardless of whether data were missing on variable *Y*. As a result, more of the data are used for each analysis than when listwise deletion is utilized, resulting in increased power.

The downsides of pairwise deletion make it difficult to use in practice. Pairwise deletion results in high model non-convergence rates, due to nonpositive definite matrices (Enders, 2001; Enders & Bandalos, 2001). Nonpositive definite matrices occur when correlation and covariance matrices are obtained using pairwise deletion that are impossible in cases of complete data. Nonpositive definite matrices often result from pairwise deletion due to different elements of the correlation and covariance matrices being computed using a different sample when using pairwise deletion. When nonpositive definite matrices occur, many common statistical models cannot be

estimated. Additionally, because the sample size varies by parameter estimate, it becomes difficult to calculate standard errors. Like listwise deletion, parameter estimates when pairwise deletion is utilized are biased when data do not meet the MCAR assumption. Thus, pairwise deletion is not a recommended technique for dealing with missing data.

Single imputation methods. Single imputation methods involve replacing missing data with calculated values based on the observed data. The majority of these techniques result in severely biased parameter estimates under all missing data conditions. However, it is useful to understand single imputation techniques, as the more useful multiple imputation (MI) technique involves many of the same concepts.

Mean imputation. Mean imputation involves replacing missing data with the mean of the observed data for that variable. Given that the missing data are replaced by the mean, mean estimates are identical to those produced by listwise deletion. However, the standard errors of mean estimates are severely attenuated under mean imputation (Olinski, Chen, & Harlow, 2003). Further, mean imputation severely attenuates estimates of variability and association between variables. Thus, mean imputation is not recommended under any missing data situations.

Regression imputation. Regression imputation involves replacing missing data with the predicted values from a regression equation (Buck, 1960). The regression equation used can involve one or multiple independent variables. One approach is to use every variable for which a case has observed data in the regression equation to impute missing values for that case. This approach results in imputed data values that fall perfectly on the regression line used to impute these values. Predictably, this approach yields negatively biased estimates of variability, and positively biased estimates of

association (Beale & Little, 1975). Although corrections are available that result in unbiased estimates of association under MCAR conditions, these corrections are rarely used in current research due to better missing data techniques being available (see multiple imputation and full information maximum likelihood below). Like mean imputation, regression imputation is generally regarded as a historical artifact and is not recommended.

Stochastic regression imputation. Stochastic regression imputation modifies traditional regression imputation by adding a residual term to account for uncertainty in the regression equation. This residual term is normally distributed with a mean of 0 and a variance equal to the error variance in the regression equation. For example, imagine we are stochastically imputing posttest scientific reasoning scores (*Y*) using pretest scientific reasoning scores (*X).* The regression equation is calculated as:

$$\hat{Y} = 4.951 + .703X$$

with a residual variance of 8.399. The intercept value of 4.951 is interpreted as the predicted posttest scientific reasoning score for a student scoring 0 at pretest. The slope value of .703 is interpreted as the increase in predicted posttest score for every unit increase in pretest score. Finally, the residual variance of 8.399 is the amount of variance in Y that is unexplained by X. In this example, the missing values of Y would be imputed by 4.951 + .703(*X*) + *e*, where *e* is a random number from a normal distribution of mean 0 and variance of 8.399. In this manner, stochastic regression reintroduces the error that is lost in traditional regression imputation. Like traditional regression imputation, computed values can also be calculated using multiple variables from the dataset in the regression equation, rather than a single variable.

Stochastic regression imputation results in unbiased parameter estimates under MCAR and MAR data mechanisms (Enders, 2010; Gold & Bentler, 2000). However, standard errors are attenuated due to the single imputation of the Y score. Single imputation techniques treat the imputed data as observed. Thus, when estimating parameter estimates from this "observed data", the certainty of the parameter estimates is overestimated, leading to underestimated standard errors. Multiple imputation (described under Modern Methods) corrects for this bias by incorporating the uncertainty involved in single imputation techniques. Given that multiple imputation is available in many software programs and stochastic regression imputation results in attenuated standard errors, stochastic regression imputation is generally not recommended over other missing data techniques. However, as noted by Enders (2010), stochastic regression imputation is involved in multiple imputation techniques.

Other single imputation methods. There are a number of other single imputation techniques that are not considered here, as they are often used in settings outside the scope of this research. These include hot-deck imputation (Ford, 1983), similar response pattern imputation (Jöreskog & Sörbom, 1993), and prorated scale scores (Keel, Mitchell, Davis, & Crow, 2002). Many of these methods result in biased parameter estimates, and all of these methods result in attenuated standard errors. Given that multiple imputation (MI) and full information maximum likelihood estimation (FIML) are readily available and do not result in biased standard errors under MCAR and MAR conditions, all single imputation techniques should be avoided in the majority of missing data situations.

Modern methods. Many of the previously reviewed methods require strict assumptions (e.g., meeting the MCAR assumption) and can result in reduced power or

biased parameter estimates. Thus, missing data methodologists almost universally recommend utilizing more modern missing data techniques when missingness is non-negligible (Allison, 2002; Enders, 2010; Little & Rubin, 2002; Schafer, 1997). Specifically, multiple imputation (MI) and full information maximum likelihood (FIML) estimation are commonly recommended. Both of these techniques result in unbiased parameter estimates and standard errors under both MCAR and MAR conditions. Further research has explored possible analytic strategies for MNAR data. However, many of these techniques require strict assumptions to be met or the researcher to specify a number of parameters a priori. Given these limitations, MNAR models are not recommended in the majority of missing data scenarios (Allison, 2002; Demirtas & Schafer, 2003; Enders, 2010; Schafer & Graham, 2002)

Multiple imputation (MI). Multiple imputation (MI) is one recommended method to deal with missing data in the majority of missing data situations. MI involves conducting multiple stochastic regression imputations, then incorporating the variability in parameter estimates across the imputations into the standard error estimates. MI is accomplished in three phases (Enders, 2010). In the *imputation phase*, multiple datasets are imputed, usually by using the data augmentation algorithm (Schafer, 1997; Tanner & Wong, 1987). In the *analysis phase*, parameter estimates are calculated for each imputed dataset separately. In the *pooling phase*, these parameter estimates are combined to produce unbiased parameter and standard error estimates. Each of these phases are outlined below.

Imputation phase. The imputation phase makes heavy use of the Bayesian framework to create multiple imputed datasets (Rubin, 1987; Enders, 2010). Utilizing the

data augmentation algorithm consists of two steps that repeat in an iterative fashion: the *imputation step,* or *I-step,* and the *posterior step,* or *P-step*.

The *I-step* involves using stochastic regression to impute the missing values. For the initial I-step, the stochastic regression coefficients are obtained using the mean vector and covariance matrix elements estimated using the available data (i.e., pairwise deletion for each of the parameter estimates). All variables included in the imputation process are used to create the stochastic regression equation for the variable with missing values. In the previous pre-post scientific reasoning example, pretest scores would be used in the stochastic regression equation to predict posttest scores. Auxiliary variables can also be included to improve the imputation of the variable with missing data.

The *P-step* involves using the dataset generated during the I-step to estimate new mean vector and covariance matrix elements. In a Bayesian framework, these elements are conceptualized as random variables with their own posterior distributions. In the P-step, new mean vector and covariance matrix elements are randomly selected from their respective posterior distributions, which are estimated using the imputed values from the previous I-step. The I-step is then repeated, using the newly-estimated mean vector and covariance matrix elements to re-estimate the stochastic regression imputation parameters and impute new values for the missing data. Thus, every I-step that is executed creates a new imputed dataset. The I-steps and P-steps can be repeated indefinitely, to create an infinite number of imputed datasets. This chain of successive I- and P-steps is considered a type of Markov Chain Monte Carlo (MCMC) procedure (Jackman, 2000).

There are two important decisions that must be made by the researcher during the imputation phase. First, the number of iterations (i.e., number of successive I- and P-

steps) needed to reach convergence must be determined. With this procedure, convergence is achieved when the posterior distributions of the mean vector and covariance matrix elements are stable. Enders (2010) recommends assessing convergence through visual analysis of *time series plots* and *autocorrelation function plots*. Time series plots display the estimated mean vector and covariance matrix elements for each successive iteration. The researcher should assess these plots for patterns, and note the number of iterations at which the plots show repeating patterns. The number of iterations at which these plots show repeated patterns indicates convergence. Autocorrelation function plots quantify the dependency between successive iterations, and can indicate the number of iterations needed between imputed datasets to ensure that parameter values are independent. Gelman and Rubin (1992) also recommend examining proportional scale reduction (PSR) values. PSR values quantify the average ratio of parameter values between two MCMC chains. If the posterior distributions for the estimated parameters are similar and stable at a given number of iterations for both chains, then PSR values will approach 1. The default convergence criteria in Mplus Version 7.11 (Muthén & Muthén, 1998-2013) is a PSR < 1.05, but stricter criteria may be applied.

The researcher should combine information from time series plots, autocorrelation plots, and PSR values to determine the number of iterations needed between each imputed dataset. These plots and values should also be assessed using multiple starting values, to ensure that one MCMC chain was not simply aberrant. Specifying too few iterations can result in correlated imputations and negatively biased standard errors, but specifying too many iterations is not problematic (Enders, 2010). Thus, the maximum

number of iterations suggested by time series plots, autocorrelation plots, and/or PSR values should generally be used between imputed datasets.

After determining the number of iterations between each imputed dataset, the researcher must determine the number of imputed datasets that will be retained for the analysis and pooling phases. Although early research suggested only three to five imputed datasets (Rubin, 1987, 1996; Schafer, 1997; Schafer & Olsen, 1998), recent research indicates that more imputations are needed to accurately estimate standard errors and maximize power (Graham, Olchowski, & Gilreath, 2007). Even at high proportions of missingness, 20 imputations have been found to give accurate standard errors. Thus, a minimum of 20 imputations is generally recommended for the majority of analyses (Enders, 2010).

Analysis phase. After imputing multiple datasets, the analysis phase involves conducting the desired analysis for each imputed dataset. In the pre-post scientific reasoning example, the mean difference between pretest and posttest scores would be computed for each imputed dataset, along with the standard error associated with this parameter. The analysis phase can be done manually for each imputed dataset, although Mplus (Muthén & Muthén, 1998-2013) and other software packages include utilities that automatically conduct the same analysis for all imputed datasets. Parameter and standard error estimates derived during the analysis phase will then be combined in the pooling phase.

Pooling phase. The pooling phase involves combining the parameter estimates and standard errors obtained for each imputation in the analysis phase. The combined parameter estimates are simply the arithmetic means of the parameter estimates obtained

for each individual imputation. For the pre-post scientific reasoning example, the pooling phase would involve computing the mean of the mean difference estimates across all imputations.

Pooling the standard errors across imputations involves combining the within-imputation parameter variance with the between-imputation parameter variance, by (from Enders, 2010, p. 223):

$$V_T = V_W + V_B + \frac{V_B}{m} \tag{1}$$

where V_T is the total sampling variance associated with a parameter, V_W is the average within-imputation parameter variance, V_B is the between-imputation parameter variance, and m is the number of imputations. V_W is calculated as (from Enders, 2010, p. 222):

$$V_W = \frac{1}{m}\sum_{t=1}^{m} SE_t^2 \tag{2}$$

simply taking the average of the squared standard errors across all imputations. V_B is calculated as (from Enders, 2010, p. 222):

$$V_B = \frac{1}{m-1}\sum_{t=1}^{m} (\hat{\theta}_t - \bar{\theta})^2 \tag{3}$$

computing the variance of individual parameter estimates $\hat{\theta}_t$ across imputations. One may notice that these individual parameter estimates also have standard errors associated with them that is not quantified in Equation 3. Thus, the V_B / m term is included in Equation 1 to account for this uncertainty. Taking the square-root of V_T gives the pooled standard error.

As mentioned previously, stochastic regression imputation (which MI is largely based on) produces unbiased parameter estimates under a MAR mechanism, but

negatively biased standard errors (Gold & Bentler, 2000). By combining between- and within-imputation error in the computation of pooled standard errors, MI corrects for this bias and produces unbiased standard error estimates when the MAR assumption is met. MI also produces unbiased parameter and standard error estimates under MCAR conditions, but results in biased parameter estimates under MNAR conditions. As will be discussed later, the accuracy of MI can be improved by the inclusion of auxiliary variables that aid in the imputation phase of the multiple imputation process.

Full information maximum likelihood (FIML) estimation. A viable alternative to multiple imputation for a researcher wanting to account for missing data appropriately is estimating model parameters using full information maximum likelihood (FIML) estimation (Enders, 2010; Schafer & Graham, 2002). Generally, maximum likelihood (ML) estimation uses an iterative procedure to determine the parameters most likely to give rise to the observed data. Many software programs, such as Mplus (Muthén & Muthén, 1998-2012), offer the option to utilize limited-information ML estimation or FIML for many analyses. Limited information ML analyzes a covariance matrix and mean vector, whereas FIML analyzes the observed data. When missingness is present, this covariance matrix and mean vector are computed using available data (i.e., pairwise deletion), and are thus only accurate under MCAR conditions. By utilizing FIML, cases with missing data are retained and their data are used in the estimation of parameters and standard errors.

FIML estimates the population parameter values that maximize the average log-likelihood of the observed data. For a single, complete case, the log-likelihood would be computed as (from Enders, 2010, p. 88):

$$\log L_i = -\frac{k}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{Y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}) \quad (4)$$

where $\log L_i$ is the log-likelihood associated with case *i*, *k* is the number of variables, $\boldsymbol{\Sigma}$ is the estimated population covariance matrix, $\boldsymbol{\mu}$ is the estimated population mean vector, and $\mathbf{Y}_i$ is the score vector for case *i*. The individual log-likelihood values quantify the relative probability of an individual's data in a multivariate normal population distribution, given a particular mean vector and covariance matrix. The individual log-likelihood value for a case with missing data is slightly modified (from Enders, 2010, p. 88):

$$\log L_i = -\frac{k_i}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\mathbf{Y}_i - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i) \quad (5)$$

with *i* subscripts associated with the number of variables, the covariance matrix, and the mean vector. The *i* subscripts indicate that these elements are allowed to vary by individual case, dependent on the variables that are missing. That is, missing variables are not included in the computation of an individual's log-likelihood value.

Like MI, FIML results in unbiased parameter estimates and standard errors under MCAR and MAR conditions, but results in bias under MNAR conditions (Enders, 2010; Little & Rubin, 2002). MI and FIML analyses tend to produce similar results if the imputation model and the maximum likelihood analysis model are *congenial* (Meng, 1994). However, the results obtained with these two techniques can differ under some circumstances. Recall that the researcher specifies the variables to be used in the multiple imputation process to help predict the variable with missingness in the stochastic regression equations used in the I-steps. If the set of variables included in the FIML analysis model differ from the set of variables included in the MI model, results will

differ across the two techniques. Additionally, the MI procedure allows all variables to relate directly to the variable with missingness. If the FIML analysis model is constrained in a way that does not allow for direct relationships between the set of variables and the variable with missingness, then the MI and FIML models will be uncongenial. FIML accuracy will be reduced if the variables with constraints are important predictors of the variable with missingness. In the pre-post scientific reasoning example with posttest missingness, suppose that a researcher specified a model where self-efficacy completely mediated the relationship between pretest and posttest scientific reasoning scores. That is, the researcher specifies a model where pretest scores do not have a direct influence on posttest scores, but rather influence posttest scores through self-efficacy. If, in reality, pretest scores have a direct effect on posttest scores, the parameter estimates associated with posttest scores would be biased. In this situation, the FIML analysis model would be misspecified, in that parameters are constrained to implausible values. Thus, for FIML results to be comparable to MI results, the FIML analysis model should be correctly specified and include all variables included in the imputation model (Collins, Schafer, & Kam, 2001; Enders, 2010; Schafer, 2003).

Utilizing auxiliary variables with MI and FIML. The accuracy of both MI and FIML results can be improved by the inclusion of auxiliary variables. Auxiliary variables are not of central interest to the substantive research questions, but are included due to their relationship with missingness (*R*) or with the variable with missing values (*Y*). Specifically, the inclusion of auxiliary variables can determine whether the MAR assumption is satisfied (Enders, 2010; Rubin, 1976). Recall the MAR mechanism requires that missingness (*R)* and the variable with missing data (*Y*) are unrelated, after

controlling for other variables in the analysis. In the pre-post scientific reasoning example, assume that students missing at posttest failed to attend the testing session due to low self-efficacy. These students with low self-efficacy would have scored lower on the scientific reasoning test than the students actually completing the scientific reasoning posttest. In this case, missingness (*R*) is related to posttest score (*Y*), but this relationship is due to self-efficacy (*X*). Although self-efficacy is not of direct interest to the assessment practitioner, it should be measured and included as an auxiliary variable. If self-efficacy is included in the MI or FIML model, then missingness is no longer related to posttest score after controlling for self-efficacy. Thus, after the inclusion of self-efficacy in the MI or FIML model, the missingness would satisfy the MAR assumption, and estimates of pre-post growth should be accurate. However, if self-efficacy is *not* included in the MI or FIML model, missingness remains related to posttest score after controlling for the included variables in the analysis, and the missingness data mechanism should be considered MNAR. In this case, the estimates of pre-post growth would be biased. Thus, it is important to include all relevant auxiliary variables in MI or FIML analyses.

Which auxiliary variables should be included? Other than estimation and computational difficulties, there is little downside to implementing an inclusive analysis strategy by including all relevant auxiliary variables. Collins and colleagues (2001) differentiated between three categories of auxiliary variables. The category of auxiliary variable depends on the variable's relationships with missingness (*R*) and with the variable with missingness (*Y*). Category A variables correlate with both *R* and *Y*, category B variables correlate with *Y* only, and category C variables correlate with *R* only. Collins

and colleagues (2001) investigated the impact of the inclusion of these auxiliary variables in parameter and standard error bias. The inclusion of category A variables in MI or FIML analyses was found to substantially reduce parameter bias and standard errors. In particular, the exclusion of a category A variable that was linearly related to *R* substantially biased mean estimates for variable *Y*, even at small (25%) proportions of missingness. Further, the variance and covariance estimates associated with variable *Y* were also biased in this case. The exclusion of a category A variable that was nonlinearly related to *R* biased variance and covariance estimates associated with variable *Y*, but not mean estimates. Under both MAR and MCAR conditions, the inclusion of category B variables reduced standard errors. The inclusion of some category C variables had no effect on parameter bias or standard errors, but the inclusion of a large number (25-50) of category C variables resulted in substantial variance and covariance estimate bias.

A close examination of the results obtained by Collins and colleagues (2001) can provide guidance on the best auxiliary variables to include when conducting MI or FIML analyses with missing data. Both category A variables and category B variables were found to be beneficial in reducing standard errors. Thus, any variables (*X*) that relate to the variable with missingness (*Y*) should be included as auxiliary variables, regardless of whether these variables relate to missingness (*R*). Category C variables, which relate only to *R* and not to *Y*, were not beneficial. However, the inclusion of category C auxiliary variables was also not harmful, unless a very large number of them were included. Also, in applied research, values of *Y* will be missing for all cases where $R = 0$. An auxiliary variable may not be related to *Y* when this relationship is estimated using only cases without missingness, but may be related to *Y* if the missing data were actually observed.

Thus, the applied researcher may believe the auxiliary variable is a category C variable and should be excluded, when it is actually a category B variable and should be included.

Due to the difficulty in accurately estimating the relationship between auxiliary variables (*X*) and the variable with missingness (*Y*) prior to conducting MI or FIML analyses, missing data experts have generally recommended an inclusive analysis strategy regarding auxiliary variables (Collins et al., 2001; Enders, 2010; Schafer, 1997). Using this strategy, any variable (*X*) with a significant relationship with missingness (*R*) or the variable with missing values (*Y*) should be included as an auxiliary variable. Although this strategy may result in the inclusion of some category C variables, the potential bias and power reduction associated with including too many category C variables is outweighed by the bias and power reduction associated with excluding category A or B variables. It should be noted, however, that some previous research indicates that including auxiliary variables with weak relationships to variables with missingness (with correlations ranging from .1 to .3) may actually reduce power when conducting FIML analyses (Savalei & Bentler, 2009). Thus, although an inclusive analysis strategy is generally recommended, it is unclear whether the inclusion of many different auxiliary variables is always beneficial.

Recent research has challenged these inclusive analysis recommendations in some special cases. Specifically, Thoemmes and Rose (in press) noted that conditioning on some auxiliary variables may lead to an *increased* conditional relationship between missingness (*R*) and missing values (*Y*). In this case, mean estimates will be *more biased* if this auxiliary variable is included in the analysis. For this reason, Thoemmes and Rose

(in press) labeled these variables "bias-inducing" auxiliary variable. Thus, the inclusive analysis strategy can backfire in special cases.

Specifying auxiliary variables when conducting MI. When conducting MI using the data augmentation algorithm, auxiliary variables are included in the stochastic regression equations used in the I-steps of the imputation phase. If the included auxiliary variables are significantly related to the variable with missingness, the inclusion of these variables in the imputation process should improve the prediction of the missing values, thus reducing bias and improving power. The auxiliary variables are only utilized in the imputation phase, and the imputed values are analyzed and pooled as before. Software programs such as Mplus (Muthén & Muthén, 1998-2012) allow for the easy inclusion of auxiliary variables in the imputation process.

Specifying auxiliary variables when conducting FIML-based analyses. Including auxiliary variables in FIML analyses involves specifying relationships with the auxiliary variables in the analysis model. Graham (2003) recommends including these variables via a *saturated correlates model*. This model is displayed graphically in Figure 3. The specification of a saturated correlates model involves allowing the auxiliary variables to correlate with explanatory variables (e.g., Pretest score in Figure 3), other auxiliary variables, and the residual terms of outcome variables. In this study, the parameters being examined are posttest mean, posttest variance, pretest-posttest covariance, and pre-post mean differences. To specify this model with auxiliary variables, pretest and posttest scores are allowed to correlate, and both of these variables are then allowed to correlate with auxiliary variables.

Auxiliary variable missingness. Just as variables of interest to the researcher can have missing values, auxiliary variables also often have missing values. Fortunately, Enders (2008) found that including important auxiliary variables with as high as 50% missingness was still beneficial in the estimation of model parameters and standard errors. Although the utility of including auxiliary variables with missingness declined as auxiliary missingness increased, particularly when the auxiliary variable was missing concurrently with the analysis variable with missing values, including an auxiliary variable with missing values was rarely harmful to the estimation of model parameters or standard errors. Thus, it is recommended to incorporate auxiliary variables with missing values into MI and FIML analyses, although these variables are somewhat less useful than auxiliary variables with complete data (Enders, 2008).

Fortunately, auxiliary variables with missingness can easily be incorporated into both MI and FIML analyses. When completing the imputation step of the MI procedure, auxiliary missing values are imputed along with the missing values of the variables of interest to the researcher. When conducting FIML analyses, auxiliary variables with missingness are included in a saturated correlates model as normal.

Methods for missing not at random (MNAR) data. In addition to the methods outlined above, there have been a number of methods proposed for missing-not-at-random (MNAR) data. The *selection model* approach (Heckman, 1976, 1979) was designed for regression models with missingness on an outcome variable. This approach involves estimating a separate regression model to predict missingness variable *R* on the outcome variable *Y*. The regression models associated with both *R* and *Y* are combined

into a path model, and the residual variance terms associated with *R* and *Y* are allowed to correlate. It is assumed that these two residual terms are bivariately normally distributed.

The *pattern mixture model* was designed for multi-wave longitudinal data with many different missing data patterns (Little, 1993). With the pattern mixture model, parameters are estimated separately for each missing data pattern. These models are underidentified, so some parameters must be fixed by the researcher to estimate the model. Commonly, the parameters associated with one of the missing data patterns are constrained to the parameters associated with the complete data.

Unfortunately, both of these models require untenable assumptions. The selection model requires strict bivariate normality of the residual terms associated with missingness *R* and outcome *Y* (Enders, 2010). The pattern mixture model requires the researcher to specify certain parameters correctly for the model to be identified. Unfortunately, neither of these assumptions is testable, and violations of these assumptions can result in significant bias (Enders, 2010; Demirtas & Schafer, 2003). Thus, MNAR models are generally not recommended, and inclusive MI and FIML analyses are considered the current state of the art (Schafer & Graham, 2002).

Missing data prevention and recovery. The previous methods have focused on various ways of analyzing data when missingness has occurred. However, preventing or recovering missing data may be the best option available to applied researchers, particularly if the mechanism underlying missingness is MNAR. A variety of strategies exist to recover data from those that drop out of a longitudinal study, such as telephoning nonrespondents (Hansen & Hurwitz, 1946) or offering an additional cash incentive to elicit responses from dropouts (Crawford, Johnson, & Laird, 1993). If recovering the

initially missing data is not possible, obtaining random samples of the missing cases can help determine the missing data mechanism (Glynn et al., 1993; Graham & Donaldson, 1993). For example, the average scores on the variable with missingness (*Y*) can be compared across initially present and initially missing cases to determine whether missingness (*R*) is related to missing values (*Y*), which would violate the MCAR assumption. Further, if enough missing data are recovered, regression models can be estimated that determine if missingness (*R*) and missing values (*Y*) remain significantly related after controlling for other dataset variables (*X*), thereby violating the MAR assumption. Unfortunately, the majority of studies examining the effects of MNAR biases have involved simulated data, and may not be representative of the MNAR mechanisms encountered by applied researchers. As noted by Graham (2009):

> Many authors have recommended collecting data on a random sample of those initially missing. However, most of this has involved simulation work and not actual data collection. Carefully conducted empirical studies along the lines suggested by Glynn et al. (1993) and Graham & Donaldson (1993) to determine the actual extent of MNAR biases would be valuable, not just to the individual empirical study, but also to the study of attrition in general. (p. 573)

Given this call for research, the current study offers a significant contribution to the missing data literature. Data that would have been missing were collected via a makeup assessment session. The precise missing data mechanism (MCAR, MAR, or MNAR) can be determined, and the extent to which this missingness biases results using various missing data handling techniques can be directly assessed. Thus, as Graham (2009) notes,

the results of this research provide significant value to assessment practice and to "the study of attrition in general."

CHAPTER THREE

Methods

Participants and Procedure

Data for the current study were collected at a mid-sized, southeastern public university. As mentioned previously, for the university to assess educational effectiveness, students are required to attend two mandatory university-wide testing sessions, labeled "Assessment Days". Assessment Day tests are administered to students twice during their undergraduate careers – once in the fall before students begin classes as entering first-year students, and once in the spring after students accumulate between 45 and 70 credit hours. Fall Assessment Days are integrated into new student orientation activities. Thus, very few students fail to attend Fall Assessment Day. When students accumulate between 45 and 70 credit hours before the beginning of a Spring semester, they are notified via email that they are required to attend the Spring Assessment Day. Despite these Assessment Day sessions being university requirements, a number of students fail to attend the posttest testing session, and these students are compelled to attend a makeup testing session to be able to register for next semester classes. The purpose of this study was to investigate the mechanism underlying missingness due to failure to attend the second mandatory testing session, and the bias introduced by treating these students' posttest data as missing. For the current study, pretest assessment data were collected during the Assessment Day conducted in Fall 2007, and posttest assessment data were collected during the Spring 2009 Assessment Day and associated makeup testing sessions.

The population of interest to university administrators is students completing between 45 and 70 credit hours within their first three semesters attending the university. That is, growth estimates are computed utilizing students who completed between 45 and 70 credit hours within three semesters after completing the pretest, and are thus invited to complete the posttest during the Spring semester of their sophomore year. In any given year, approximately 2/3 of the students completing a given pretest are invited to complete the posttest three semesters later *due to their completion of 45 -70 credit hours during the prior three semesters*. The vast majority of the remaining 1/3 of the pretest population are invited to complete the posttest five semesters after entering the university due to earning less than 15 credits per semester. Importantly, the assessment design utilized at the university only matches pretest and posttest assessment data for students completing posttest *three semesters after completing pretest* (i.e., the university only computes "value-added" estimates for this specific population of interest). Thus, students completing assessments after five semesters of university attendance are not considered the population of interest by the university. Given the university utilizes Assessment Day data to measure the impact of the first three semesters of university attendance, only students who 1) completed the pretest as entering freshmen in Fall 2007 *and* 2) earned 45-70 university credits during their first three semesters at the university, which resulted in a requirement to complete the posttest during the Spring 2009 semester, were examined in this study.

Noncognitive test sample. All 3,766 incoming first-year students completed a three-item noncognitive measure of mastery orientation towards learning (Mastery Approach, Achievement Goal Questionnaire; Finney, Pieper, & Barron, 2004) during the

Fall 2007 Assessment Day. Mastery orientation scores were not available for the 78 students attending a *pretest* makeup session. Given the low number of pretest makeup students, this study only focused on the effects of *posttest* nonattendance on growth estimates. Of the 3,766 students completing pretest, 2,321 students completed between 45 and 70 credit hours within the first three semesters of university attendance and completed the posttest in Spring 2009. Note that students that fail to complete any of the three mastery orientation items are not given a total score on mastery orientation. Of the original sample of 2,321 students, 67 students (63 Assessment Day attendees and 4 makeup attendees) did not provide complete item responses at pretest or posttest. Recall that the purpose of this study is to examine the impact of missingness due to posttest nonattendance. Although these 67 students have missing data, this missingness was not examined in this study. Thus, these 67 students were excluded from further analysis.

Of the remaining 2,254 students invited to attend the Spring 2009 Assessment Day to complete the posttest, a subset of 2,120 students (94.1%) attended Assessment Day, whereas 134 (5.9%) skipped Assessment Day (i.e., initially missing) and subsequently attended a makeup assessment session. The 2,120 students representing the "Assessment Day" sample were 65.2% female, 84.1% White, 4.7% Asian, 2.6% Black, 2.8% Hispanic, 0.5% Pacific Islander, and 5.2% unspecified ethnic origin. This sample had an average age of 19.92 years (SD = 0.37) at posttest. The 134 students representing the "Makeup" sample were 48.5% female, 80.6% White, 2.2% Asian, 3.0% Black, 3.0% Hispanic, 0.8% Pacific Islander, and 10.5% unspecified ethnic origin. This sample had an average age of 19.97 years (SD = 0.48) at posttest.

A cursory comparison of the demographic information for the Assessment Day and makeup samples indicated that a MCAR may not underlie the makeup noncognitive test data. The proportion of males that attended the makeup testing sessions was greater than the proportion of males that attended Assessment Day. If this difference was statistically significant, the "missingness" being investigated (i.e., whether a student attends Assessment Day or a makeup session at posttest) would be significantly related to an observed variable in the dataset (gender), thus ruling out a MCAR mechanism. This difference in proportions and other associations between dataset variables and posttest attendance were examined when screening for potential auxiliary variables (described later in Chapter 3).

Cognitive test sample. A random sample of 1,486 incoming first-year students completed a 66-item cognitive test of scientific reasoning (Natural World, Version 9, Sundre, 2008) during the Fall 2007 Assessment Day. Note that this number includes 78 students who attended a makeup testing session at *pretest.* Of the 1,486 students completing the scientific reasoning pretest, 835 students completed between 45 and 70 credit hours within the first three semesters of university attendance and thus completed this same test in Spring 2009 (posttest). Of the 835 students, 789 students (94.5%) attended their assigned Assessment Day testing session, whereas 46 students (5.5%) were compelled to attend a makeup assessment session. The 789 students attending "Assessment Day" were 65.5% female, 85.4% White, 4.3% Asian, 2.9% Black, 2.4% Hispanic, 0.1% Pacific Islander, and 4.7% unspecified ethnic origin. This sample had an average age of 19.93 years (SD = 0.37) at posttest. The 46 students skipping Assessing Day (i.e., initially missing) and later attending a makeup testing session were 43.5%

female, 82.6% White, 2.2% Black, and 15.2% unspecified ethnic origin. This sample had an average age of 19.92 years (*SD* = 0.31) at posttest. When scoring the scientific reasoning test, unanswered items are marked as incorrect. A total score was obtained for all 835 students at both pretest and posttest.

Similar to the demographic information obtained for the noncognitive test sample, the proportion of makeup students that were male was greater than the proportion of Assessment Day attendees that were male. Again, if this difference is statistically significant, "missingness" (i.e., whether a student attends Assessment Day or a makeup session at posttest) would be significantly related to an observed variable in the dataset (gender), thus ruling out a MCAR mechanism.

Noncognitive accountability measure – Mastery Approach (MAP) Goal Orientation. The Mastery Approach Goal Orientation Subscale (MAP) of the Achievement Goal Questionnaire (AGQ; Finney et al., 2004) is a three-item measure of the extent to which a student is motivated to master course material with the goal of developing competence. Examinees respond to MAP statements on a Likert scale from 1 ("Not at all true of me") to 7 ("Very true of me"). Total scores were computed by summing the scores to the three items, and thus can range from 3 to 21. Previous research has found MAP scores to be relatively reliable, with Cronbach's coefficient alpha estimates typically ranging between .70 and .80. In the current study, MAP alpha estimates were .75, .81, and .82 for the pretest, posttest Assessment Day, and posttest makeup administrations, respectively.

Cognitive accountability measure – Natural World Version 9. The Natural World, Version 9 test (NW-9; Sundre, 2008) is a 66-item cognitive test designed to

measure quantitative and scientific reasoning skills. Items are scored correct or incorrect, and summed to create one total scientific reasoning score. Items were designed by a team of mathematics and science faculty members working in conjunction with assessment and measurement experts. In previous samples, NW-9 scores have been fairly reliable, with Cronbach's coefficient alpha estimates typically ranging between .70 and .90. In the current study, NW-9 alpha estimates were .79, .81, and .87 for the pretest, posttest Assessment Day, and posttest makeup administrations, respectively.

Auxiliary variables. Auxiliary variables were used in two ways in this study. First, auxiliary variables were used to help identify the missing data mechanism. That is, missingness (*R*, whether a student completed the posttest during Assessment Day or a makeup testing session) and posttest scores (*Y*) may be related when auxiliary variables (*X*) are excluded (i.e., data would be considered MNAR), but may *not* be related after controlling for certain auxiliary variables (i.e., data would be considered MAR when including auxiliary variables). Thus, examining the relationships between missingness (*R*), posttest scores (*Y*), and other dataset variables (*X*s) was important to fully understand the missing data mechanism.

Second, after identifying the missing data mechanism, auxiliary variables were integrated into the MI imputation model and the FIML analysis model to determine the effects when including these variables. The inclusion of important auxiliary variables was expected to influence parameter estimates and standard errors in a manner consistent with prior research (Collins et al., 2001). Table 2 summarizes these expectations. Thus, the choice of auxiliary variables was thoughtful to 1) identify the missing data mechanism and maximize the probability of meeting the MAR assumption, and 2) to evaluate the

impact of including quality auxiliary variables on both parameter estimates and standard errors.

As mentioned previously, it is generally recommended that auxiliary variables be included in the analysis if they are significantly related to either missingness or to the variable with missing values (Enders, 2010). For the purposes of this study, it was also important to consider the auxiliary variables that would be available in a typical operational testing program. For example, scores on the same variable measured at multiple time points are often very highly correlated (Raymond, Neustel, & Anderson, 2009). Thus, pretest score is recommended to be included as an auxiliary variable when imputing/analyzing posttest scores (Graham, 2009). Pretest scores are often readily available in the context of higher education accountability data. That is, the university typically collects pretest data as a part of the assessment design.

Additionally, a number of variables are commonly available to university assessment coordinators through university student information systems. These variables often include general demographic information (e.g., gender, age), admission test scores (e.g., SAT scores), as well as college performance and completion measures (e.g., GPA, credit hours completed). Given their ready availability at many institutions, these variables would be ideal candidates for auxiliary variables.

The current assessment design involves measuring a number of other constructs besides mastery orientation and scientific reasoning at both pretest and posttest. However, note that the typical assessment practitioner would *not* have access to posttest auxiliary variable scores if makeup data were not collected. That is, if a practitioner were to utilize MI or FIML *instead of* collecting makeup assessment data, that practitioner

would *not* be able to utilize *posttest* auxiliary variables, because scores on these variables would not be collected. However, this practitioner would have access to *pretest* auxiliary variable scores, which may serve as sufficient proxies of posttest auxiliary variable scores for the purposes of imputing and/or analyzing posttest scientific reasoning or mastery orientation scores with missingness. Thus, this study examined the utility of pretest auxiliary variable scores as proxies of posttest auxiliary variable scores, as detailed under Research Question 2 in Chapter 3.

Ideally, only the most accessible auxiliary variables would be needed to meet the MAR assumption. For example, pretest scientific reasoning scores are likely to be highly related to posttest scientific reasoning scores, and they are already measured as a part of the typical assessment design. Even if missingness (R) was related to posttest scores (Y), this relationship may no longer be significant after controlling for pretest scores (X), thus meeting the MAR assumption. In this case, the auxiliary variables that are more difficult to obtain would be unnecessary to meet the MAR assumption. As detailed later in this chapter under Research Question 2, this study examined the extent to which the missingness mechanism would be considered MAR or MNAR after including different sets of auxiliary variables. Thus, practitioners could use this information to determine which variables would need to be obtained and included as auxiliary variables to obtain accurate parameter estimates and reduce standard errors.

Auxiliary variables hypothesized to be related to missingness. Students attending the makeup testing sessions at posttest have been found to differ from students attending the Assessment Day testing sessions in a number of ways. Given that "missingness" is operationalized as attending a makeup testing session in this study, the variables with

differences between Assessment Day and makeup samples can potentially be utilized as auxiliary variables. Swerdzewski and colleagues (2009) found makeup students are more likely to be men, with makeup sessions comprised of 46% male students, as opposed to 36% male students during the typical Assessment Day sessions. Compared to students attending Assessment Day, makeup students were also found to be older (d = .36), have lower GPAs (d = -.39), and have a higher number of earned credits at posttest (d = .28) than students attending Assessment Day. Additionally, compared to students attending the Assessment Day testing sessions, makeup students have also been found to have lower MAP scores (d = -.32), lower scores on a measure of performance-approach goal orientation (PAP, the motive to perform better than other students; d = .27), higher scores on work avoidance related to coursework (WAV; d = -.35), lower conscientiousness scores (d = -.28; Zilberberg, 2013), and also report lower test-taking effort (d = -.42) and perceived test importance (d = -.25).

Combining the information from previous research creates a profile of the typical student attending a makeup testing session. This typical makeup student is more likely to be male, older, have a lower GPA, and have higher earned credits. The examinee also tends to be less motivated to perform well academically, less conscientious, more work avoidant, and less willing to put forth effort on tests or find them important. However, it was unclear if these relationships would replicate with the sample being used in the current study. It was also unclear whether all of the variables defining this student profile would also relate to posttest mastery orientation or scientific reasoning scores. Any variables hypothesized to relate to both Assessment Day attendance *and* posttest mastery orientation or scientific reasoning scores would be ideal candidates for auxiliary

variables, as including these variables would both reduce parameter bias and standard errors (Collins et al., 2001). However, if the variables defining the makeup student profile were *not* related to posttest scores, then including these variables as auxiliary variables in MI or FIML analyses is not likely to aid in parameter estimation. Thus, it was important to also examine variables that have been found to relate to mastery orientation or scientific reasoning scores.

Given the relationships to "missingness" (i.e., Assessment Day vs. makeup attendance) discovered in previous research, gender, posttest age, posttest GPA, and total credits completed at posttest were obtained from the university student database and utilized as auxiliary variables for both mastery approach and scientific reasoning growth analyses (see Table 4). Additionally, both pretest and posttest scores on PAP, WAV, conscientiousness, and test-taking effort and importance were utilized as auxiliary variables for both mastery approach and scientific reasoning growth analyses. MAP pretest and posttest scores were used as auxiliary variables for scientific reasoning growth analyses. Note that pretest MAP scores are automatically included in the MI imputation model and FIML analysis model when conducting the mastery orientation growth analyses.

Auxiliary variables hypothesized to be related to MAP scores. Previous research has found gender and SAT scores to predict MAP scores (Davis, Pastor, & Barron, 2004). Across multiple studies, MAP scores have been found to positively relate to performance-approach (PAP) scores (r = .28-.42), mastery-avoidance (MAV, the motive to avoid learning less than possible) scores (r = .22-.27), performance-avoidance (PAV, the motive to avoid performing worse than other students) scores (r = .06-.13) and work

avoidance (WAV) scores (r = -.58; Finney et al., 2004; Pieper, 2003). Additionally, mastery-approach orientation has been found to relate to Big Five personality variables, positively correlating with Openness (r = .44), Conscientiousness (r = .32), Extraversion (r = .29), and Agreeableness (r = .19), and negatively correlating with Neuroticism (r = -.18; Payne, Youngcourt, & Beaubien, 2007). Mastery-approach orientation has also been found to relate positively to metacognitive strategies (r = .48; Howell & Watson, 2007), as well as test-taking effort and perceived test importance (effort r = .27-.34, importance r = .09-.23; Barry, 2010).

Given this previous research, gender, SAT Math and Verbal scores, pretest metacognitive regulation scores, and pretest and posttest PAP, MAV, PAV, WAV, Big Five, and test-taking effort and importance scores were assessed as possible auxiliary variables. Unfortunately, metacognitive regulation was not measured at posttest due to testing time constraints. Fortunately, many of these variables were hypothesized to relate to *both* posttest mastery orientation scores and Assessment Day attendance (see Table 4). If these variables were found to relate to both "missingness" (Assessment Day vs. makeup) and posttest mastery orientation scores, the inclusion of these auxiliary variables in MI and/or FIML analyses should reduce both standard errors and parameter bias associated with posttest mastery orientation scores.

Auxiliary variables hypothesized to be related to NW-9 performance. Previous research has found the number of science credits completed by a student at posttest to be predictive of NW-9 scores, with students completing four or more science credits scoring five raw score points higher on average than students completing no science credits (Sundre, 2008). SAT Math scores have been found to be predictive of NW-9 test scores

(*r* = .46; Barry, 2010), and both SAT Math and SAT Verbal scores have been found to be predictive of scores on a previous version of the Natural World test (*r* = .38 and .46, respectively; Wise, Wise, & Bhola, 2006). Metacognitive regulation, or a student's ability to regulate their own learning processes, has been found to be predictive of seventh grade English and science exams (*r* = .28; Pintrich & De Groot, 1990). Combined test-taking effort and importance were also found to relate to a previous Natural World test (*r* = .33, Sundre & Wise, 2003).

Given these relations with cognitive test performance, SAT Math and Verbal scores, posttest earned science credits, pretest metacognitive regulation scores, and pretest and posttest test-taking effort and importance scores were examined as possible auxiliary variables. For the scientific reasoning scores, only test-taking effort and importance scores were hypothesized to relate to both "missingness" (Assessment Day vs. makeup) and posttest scientific reasoning scores (see Table 4). However, some variables hypothesized to be related to Assessment Day attendance have not been examined for relationships with scientific reasoning scores (e.g., age). Thus, these variables may be related to scientific reasoning scores, and thus may reduce standard errors and parameter bias when included as auxiliary variables.

Auxiliary variable measures. The aforementioned auxiliary variables are presented in Table 4, along with their missingness proportions for both the noncognitive and cognitive test samples. Missingness proportions vary across measures due to some measures only being administered in certain testing configurations. Gender, age at posttest, SAT Math and Verbal scores, posttest GPA, total earned credits, and earned science credits were obtained via the university information system. Scores on the

remaining auxiliary variables were collected by administering the instruments outlined below.

Achievement Goal Questionnaire (AGQ). The 16-item Achievement Goal Questionnaire (AGQ, Finney et al., 2004; Pieper, 2003) measures goal orientations relevant to learning and performance in college. Examinees respond to statements on a Likert scale from 1 ("Not at all true of me") to 7 ("Very true of me"). The original measure consisted of four subscales, measuring mastery-approach (MAP, motive to master course material), performance-approach (PAP, motive to perform well relative to others), mastery-avoidance (MAV, motive to avoid learning less than possible) and performance-avoidance (PAV, motive to avoid performing worse than others) goal orientations. Pieper (2003) added four additional work avoidance (WAV) items, to measure the motive to avoid doing coursework. MAP, PAP, MAV, and PAV scores can range from 3 to 21, and WAV scores can range from 4 to 28.

Big Five Inventory (BFI-44). The Big Five inventory (BFI-44, John & Srivastava, 1999) is a 44-item measure designed to assess five dimensions of personality. These five dimensions include Openness (intellectual, imaginative, independent-minded), Conscientiousness (orderly, responsible, dependable), Extraversion (talkative, assertive, energetic), Agreeableness (good-natured, cooperative, trustful), and Neuroticism (uncalm, easily upset) (John & Srivastava, 1999). Participants were asked to respond to a series of statements using a scale from 1 ("Disagree Strongly") to 5 ("Agree Strongly"). Extraversion and Neuroticism were each measured by 8 items (with scores ranging from 8 to 40), Agreeableness and Conscientiousness were each measured by 9 items (with

scores ranging from 9 to 45), and Openness was measured by 10 items (with scores ranging from 10 to 50).

Metacognitive Awareness Inventory - Regulation (MAI-R). The Regulation subscale of the Metacognitive Awareness Inventory (MAI-R; Schraw & Dennison, 1994) is a 35-item measure designed to assess the ability to implement study strategies to regulate one's learning. Participants were asked to respond to a series of statements using a scale from 1 ("Always False") to 5 ("Always True"). Thus, scores ranged from 35 to 175. This measure was only administered during the Fall 2007 pretest, and not the Spring 2009 posttest.

Student Opinion Scale (SOS). The Student Opinion Scale (SOS; Thelk, Sundre, Horst, & Finney, 2009) is a 10-item measure designed to measure examinee test-taking motivation. The SOS consists of two 5-item subscales: Effort (how much effort the examinee reports putting forth on a test) and Importance (how much importance the examinee places on a test). Participants were asked to respond to a series of statements using a scale from 1 ("Strongly Disagree") to 5 ("Strongly Agree"). Thus, both Effort and Importance scores ranged from 5 to 25.

Data Analysis

Analyses for all of the research questions below were conducted using Mplus Version 7.11 (Muthén & Muthén, 1998-2013).

Research question 1: Examining posttest response validity. A multiple-group analysis was conducted to determine the extent to which makeup examinees are providing valid responses at posttest. As mentioned previously, students attending posttest makeup sessions may be responding randomly due to reduced test-taking effort.

As a result, the *complete* dataset analyses would be biased by *including* the makeup data, as the growth estimates obtained using these data would not be representative of true student growth. To examine this possibility, multiple-group models were specified predicting posttest scores from pretest scores for both the Assessment Day and makeup samples (see Figure 4). Posttest scores were regressed on pretest scores as:

$$Y = i + bX + e \tag{6}$$

where Y is posttest score, X is pretest score, i is an intercept parameter, b is the slope predicting posttest score (Y) from pretest score (X), and e is a normally distributed residual term representing the variance in posttest score (Y) unexplained by pretest score (X). In the unconstrained model, the intercept (i), slope (b), and residual variance (e) are estimated separately for the Assessment Day and makeup samples. The fit of four constrained models were assessed to determine the extent to which students in the makeup sample provided valid responses. First, intercepts (i) were constrained to be equal across the Assessment Day and makeup samples. Second, slopes (b) were constrained to be equal across groups. Third, residual variances (e) were constrained to be equal across groups. Fourth, all regression parameters (intercepts, slopes, and residual variances) were constrained to be equal across groups. The model-data fit was examined for all four of these models. Fit was assessed by examining the χ^2 statistic, the comparative fit index (CFI), and the root mean squared error of approximation (RMSEA). A statistically significant χ^2 value indicates that the constrained model fits significantly worse than the freely-estimated model. The χ^2 statistic quantifies the absolute model-data fit, whereas the CFI and RMSEA quantify relative approximate fit (Hu & Bentler, 1998). Hu and Bentler (1999) considered CFI values larger than .95 and

RMSEA values less than .06 to indicate adequate model data fit, although Marsh, Hau, and Wen (2004) indicated that these values can be influenced by model size and variable correlation magnitude, making universal guidelines difficult to follow in practice.

In a sense, the models regression models described above are testing the assumptions made when specifying a pattern-mixture model (Little, 1993). In these models, it is assumed that a different growth pattern or relationship between variables may underlie each missing data pattern. However, these relationships are empirically underidentified, given different time points are missing for different patterns. Thus, the pattern-mixture models specify some parameters (e.g., pre-post slope) to be equivalent across missing data patterns. Given that missing data were collected during makeup sessions, these constraints can be tested for statistical and practical misfit.

Ideally, the fourth model (with all regression parameters constrained to be equivalent across groups) should sufficiently fit the data, indicating that the relationship between pretest and posttest scores remains constant across the Assessment Day and makeup samples. However, if the fourth model does *not* fit the data, this misfit could be due to less effortful responding at posttest by the makeup sample. Compared to the Assessment Day sample, this lack of effort by the makeup sample could manifest in a different intercept (e.g., makeup examinees scored lower on average on the cognitive test at posttest than would be predicted for the Assessment Day sample with the same pretest scores), lower slope (indicating that pretest scores do not predict posttest scores as strongly for makeup examinees), or increased residual variance (indicating an increase in unexplained variability in posttest scores introduced by random responding by the makeup examinees). If the fourth model does not fit the data, the first three models

should provide information on the parameters that differ across Assessment Day and makeup examinees.

Research question 2: Examining the missing data mechanism. What missing data mechanism underlies posttest non-attendance? As mentioned previously, the missing data mechanism is determined by the relationships between a dichotomous missingness variable *R*, the variable with missingness *Y*, and other dataset variables (i.e., auxiliary variables). The missing data mechanism is considered MCAR if missingness *R* is unrelated to both *Y* and other dataset variables (*X*), MAR if *R* is unrelated to *Y* conditional on other dataset variables (*X*), and MNAR if *R* remains related to *Y* conditional on other dataset variables (*X*). In most missing data scenarios, the exact mechanism cannot be determined, as values of *Y* are missing for all cases where $R = 0$. However, in this study, the initially missing posttest scores were recovered via a makeup testing session. Referring to Table 1, the values of the "missing" data (*Y*) were known, thus the missing data mechanism could be empirically determined.

To assess the linear relationship between missingness (*R*) and both the variables of interest (*Y*) and the other variables in the dataset (i.e., the auxiliary variables noted above), a series of correlation and regression models were estimated. First, to test the MCAR assumption, the simple bivariate relationships between missingness (*R*), posttest scores (*Y*), and other measured variables (*X*) were estimated. These "other measured variables" were the auxiliary variables discussed above. Given that the auxiliary variables also had missing values (see Table 4), bivariate relationships between missingness (*R*), posttest scores (*Y*), and other measured variables (*X*) were estimated following MI of all

missing auxiliary data. If missingness (*R*) was unrelated to both posttest scores (*Y*) and other measured variables (*X*), the missingness mechanism could be considered MCAR.

Second, the partial correlation between posttest attendance (*R*) and posttest scores (*Y*) was estimated after controlling for *each* of the auxiliary variables (*X*), including pretest scores. This would provide some indication of the variables that *independently* moderate the relationship between posttest attendance (*R*) and posttest scores (*Y*). If the partial correlation between posttest attendance (*R*) and posttest scores (*Y*) after controlling for a given auxiliary variable (*X*) is substantially lower than the bivariate *R*-*Y* correlation, this would indicate that the auxiliary variable (*X*) is an important moderator for the *R*-*Y* relationship, and thus should be included as an auxiliary variable to reduce parameter bias and standard errors.

Third, multiple regression analyses were conducted to further examine the missing data mechanism. Auxiliary variables (*X*s) were entered in blocks in multiple regression analyses predicting posttest scores (*Y*), in the order of their ease to obtain for the typical assessment practitioner. Pretest score on the construct of interest (scientific reasoning or mastery orientation) was entered first as the most easily accessible auxiliary variable, given pretest scores are commonly collected as part of the pre-post assessment design. Then, university student information system variables were entered, followed by pretest scores on *other* constructs (i.e., not the construct of interest), followed by posttest scores on other constructs. The variance explained (R^2) and additional variance explained by each subsequent model (R^2 change) were estimated to determine the additional predictive utility of each block of predictors. If the additional variance explained by a subsequent block of predictors was insignificant, this would indicate that the block is not

needed to predict additional posttest score variance, and thus would *not* be useful to include as auxiliary variables to reduce standard errors. This would help assessment practitioners identify the auxiliary variables that are absolutely necessary to collect to aid in MI and FIML analyses.

Fourth, the partial correlation between posttest attendance (*R*) and posttest scores (*Y*) was estimated for each of the regression models described above. The partial correlation quantifies the relationship between posttest attendance (*R*) and posttest scores (*Y*) conditional on the other variables in the regression model. If the partial correlation was negligible for a given model, the MAR assumption would be met after conditioning on the variables included within that model. However, if this partial correlation was non-negligible for a given model, then the mechanism would be considered MNAR when conditioning on the variables included within that model. Thus, examining the partial correlation values provides an indication of the circumstances under which the MAR assumption is satisfied, as well as which combination of auxiliary variables should be included in MI or FIML analyses to meet the MAR assumption.

Note that, if assessment practitioners were to forego makeup testing and instead utilize MI or FIML with auxiliary variables, they would *not* have access to auxiliary variables collected at posttest for examinees with missing posttest scores on the construct of interest (mastery orientation or scientific reasoning). However, if the MAR assumption can be met using *pretest* auxiliary variables, then posttest scores on these same auxiliary variables would be unnecessary. Note the previously referenced research established relationships between the potential auxiliary variables and Assessment Day attendance, mastery orientation scores, and/or scientific reasoning scores when these scores were

collected *at the same testing session*. That is, prior research suggests that *posttest* scores on these auxiliary variables are predictive of *posttest* Assessment Day attendance, mastery orientation, or scientific reasoning, but it is unclear whether *pretest* scores on these auxiliary variables can serve as sufficient proxies of posttest scores on these same measures. Although it is reasonable to expect pretest scores on stable constructs (e.g., conscientiousness; John & Srivastava, 1999) to serve as proxies of posttest scores on the same construct, this expectation may not hold for constructs that change substantially over time (e.g., test-taking effort, Barry, 2010). Thus, it was important to compare the impact of including pretest auxiliary variables as proxies of posttest auxiliary variables versus including the posttest auxiliary variables themselves. Assessing the utility of pretest auxiliary variables as proxies of posttest auxiliary variables involved 1) examining the bivariate correlations between pretest and posttest auxiliary variable scores to determine the stability of auxiliary variable scores over time, 2) examining the difference between how pretest versus posttest auxiliary variable scores related to posttest attendance (*R*) and posttest scores (*Y*), and 3) comparing multiple regression models including or excluding posttest auxiliary variable scores to determine if posttest auxiliary variables provided posttest score predictive utility *above and beyond* pretest auxiliary variables. If posttest auxiliary variables are only moderately correlated (i.e., not collinear) with pretest auxiliary variables, are more strongly related to posttest attendance (*R*) or posttest scores (*Y*), and/or provide additional predictive utility above and beyond pretest auxiliary variables, posttest auxiliary variables may be needed to obtain more accurate parameter estimates or standard errors from MI or FIML analyses.

Note that previous research indicates that excluding an auxiliary variable that is linearly related to the variable with missingness (*Y*) but *nonlinearly* related to missingness (*R*) can result in biased variance and covariance estimates associated with *Y* (Collins et al., 2001). Specifically, *convex* relationships, where missingness percentages are higher at the extremes of the auxiliary variable distributions, were found to result in significant variance and covariance estimate bias. Thus, overlapping density distributions of the Assessment Day and makeup samples were examined to screen for nonlinear relationships between posttest attendance (*R*) and the auxiliary variable scores (*X*). If posttest attendance (*R*) were *not* nonlinearly related to any auxiliary variables (*X*), the auxiliary variable distribution of the Assessment Day and makeup samples would have approximately equivalent shape. However, if more students from the makeup sample score in the extremes of the auxiliary variable distribution than students from the Assessment Day sample, this pattern would indicate that there is a convex relationship between posttest attendance and that auxiliary variable. If a convex relationship exists between a dataset variable and Assessment Day attendance, that dataset variable should be included as an auxiliary variable in MI and FIML analyses to reduce bias in variance and covariance estimates.

When conducting these analyses to identify the missing data mechanism, it was important to take into account both statistical significance and practical significance (i.e., effect size). For example, assume that missingness (*R*) was statistically significantly bivariately related to a dataset variable (*X*), but the point-biserial correlation between the two variables is only $r = .05$. In this case, the MCAR assumption is violated in the strict sense, but there are unlikely to be any practical consequences of this violation. That is,

utilizing listwise deletion would likely not result in large biases in posttest score (*Y*) parameters, given the practically small relationship between missingness (*R*) and posttest scores (*Y*). There is no strict cutoff for the magnitude of the relationship between *R* and *Y* that is problematic, given the parameter and standard error bias also depend on the percentage of missingness and the specific analysis being conducted (McKnight et al., 2007). However, simulation studies often create missing data by deleting values *completely dependent on* the values of the auxiliary variables (to simulate a MAR mechanism) or the values of the variable with missingness (to simulate a MNAR mechanism), creating a strong relationship between missingness (*R*) and the variable with missingness (*Y*) (e.g., Collins et al., 2001; Enders & Bandalos, 2001). Although statistical significance is mainly being considered when identifying the missing data mechanism and building auxiliary models in the current study, the magnitude of relationships between auxiliary variables (*X*), posttest scores (*Y*), and missingness (*R*) were considered when examining and interpreting the results of later analyses (see Research Question 3).

Research question 3: Comparing missing data handling techniques. To what extent are results affected by using different missing data handling techniques? Simply identifying the missing data mechanism (e.g., MCAR, MAR, MNAR) and the pattern of missingness does not indicate the extent to which results are biased by the missingness. For example, the posttest makeup assessment data could be considered MNAR, but the proportion of missingness may be low enough that MCAR- or MAR-based missing data handling techniques do not introduce practically significant bias to parameter estimates. Thus, the results of different approaches to analyzing missing data were compared to each other, and, most importantly, to the results obtained using the complete dataset.

Specifically, posttest mean, posttest variance, pretest-posttest covariance, and pre-post mean difference were estimated utilizing different missing data handling techniques and using the complete dataset. The discrepancy between the estimates and their associated standard errors obtained via different missing data techniques and the complete dataset were then examined.

Standardized parameter discrepancy was examined by:

$$sDiscrepancy = \frac{\hat{\theta}_{method} - \hat{\theta}_{complete}}{SE_{complete}} \tag{6}$$

where the parameter estimate obtained from analyzing the complete data ($\hat{\theta}_{complete}$) is subtracted from the parameter estimate obtained from utilizing a missing data handling method ($\hat{\theta}_{method}$) and divided by the standard error of the parameter estimate obtained from the analyzing the complete data ($SE_{complete}$). Standardized parameter discrepancy quantifies the standard error difference between the parameter estimate obtained by utilizing a missing data handling method and the parameter estimate obtained by analyzing the complete data. This estimate is comparable to *standardized bias* computed by Collins and colleagues (2001):

$$sBias = \frac{\hat{\theta}_{mean} - \theta}{SE} \tag{7}$$

where the average parameter estimate across replications ($\hat{\theta}_{mean}$) is subtracted from the true parameter (θ), and divided by the standard deviation of the parameter across replications (*SE*). Collins and colleagues (2001) suggest standardized bias can be interpreted similarly to Cohen's *d*, and values of > |.4| can be considered practically significant. However, the parameter estimates obtained from analyzing the complete data

($\hat{\theta}_{complete}$) and utilizing a missing data handling method ($\hat{\theta}_{method}$) are both point estimates utilizing a single sample. Thus, these estimates can be influenced substantially by sampling error. By contrast, $\hat{\theta}_{mean}$ is an *average* of parameter estimates across many replications, and is not as affected by sampling error as a single point estimates. Similarly, θ is usually set by the researcher and is assumed to be error-free. Thus, the standardized discrepancy estimates computed in this study can be substantially larger than standardized bias estimates simply due to the impact of sampling error on the parameter estimates. Standardized discrepancy, then, can be interpreted more similarly to a *z*-score rather than a Cohen's *d* estimate. For this study, standardized discrepancy values greater than |2| were considered larger than would be expected given sampling error, and were flagged as exhibiting substantial bias.

Following Arbuckle (1996) and Enders and Bandalos (2001), standard errors were compared by computing *relative efficiency* (*RE*) estimates:

$$RE = \frac{SE^2_{method}}{SE^2_{complete}} \tag{8}$$

where the squared standard error of the parameter estimate obtained via the missing data handling method (SE^2_{method}) is divided by the squared standard error of the parameter estimate obtained by analyzing the complete dataset ($SE^2_{Complete}$). Thus, values closer to 1 indicate comparable standard error estimates between the missing data handling method and the complete data, whereas values greater than 1 indicate standard error inflation due to utilizing the missing data handling method. Given that the squared standard error is inversely related to sample size, the *RE* estimate also quantifies the sample increase needed for the missing data handling method to achieve the same precision as analyzing

the complete data (Arbuckle, 1996). For example, a RE value of 1.10 would indicate that the sample size of the missing data handling method dataset would need to increase by 10% to achieve the same precision as analyzing the complete dataset.

The following are the missing data results that were compared:

Method 1 – Complete dataset: The data obtained from the makeup sample were combined with the data obtained from the Assessment Day sample. Thus, there was no missingness in this data analysis. Posttest mean, posttest variance, covariance with pretest, and pre-post mean difference estimates were then obtained using this complete dataset.

Method 2 – Listwise deletion: The makeup sample was not included in the estimation of parameters (i.e., posttest mean, posttest variance, pretest-posttest covariance, and pre-post mean difference). This method aligns with current practice associated with this large-scale testing program. Sample Mplus syntax associated with the analyses for Methods 1 and 2 is presented in Appendix A.

Method 3 – Multiple imputation without auxiliary variables: Makeup posttest data was treated as missing. The makeup posttest data values were then multiply-imputed, without utilizing any auxiliary variables. It is recommended that, at a minimum, any variables included in the analysis model should be included in the imputation model (Enders, 2010). Thus, only pretest scores were used to impute posttest scores. Given the relative efficiency of measures of association remained high in simulation studies using 20 imputations, even at high amounts of missingness (Graham, Olchowski, & Gilreath, 2007), 20 datasets were imputed. Preliminary analyses suggested that 2500-2700 iterations were sufficient for convergence across all conditions. To be conservative, 5000

iterations were used between imputed datasets in all conditions. Sample Mplus syntax specifying imputation of posttest data can be found in Appendix B, and sample analysis syntax utilizing the multiple imputed datasets can be found in Appendix D.

Method 4– Multiple imputation with university database and pretest auxiliary variables: Makeup posttest data were treated as missing. The makeup posttest data values were then multiply-imputed using university database and pretest auxiliary variables to aid in imputation. As mentioned previously, a typical assessment practitioner would not have access to posttest auxiliary variables if posttest makeup data were not collected. Thus, it was important to compare the results when including and excluding posttest auxiliary variables. As with Method 3, 20 datasets were imputed, and every 5000th iteration was extracted. Sample Mplus syntax specifying the imputation of this data can be found in Appendix C, and sample analysis syntax utilizing the multiple imputed datasets can be found in Appendix E.

Method 5 – Multiple imputation with all auxiliary variables: Makeup posttest data were treated as missing. The makeup posttest data values were then multiply-imputed using all auxiliary variables (i.e., pretest, university database, pretest auxiliary variables and posttest auxiliary variables) to aid in imputation. As with Methods 3 and 4, 20 datasets were imputed, and every 5000th iteration was extracted. Syntax for this imputation is found in Appendix D, and sample analysis syntax utilizing the multiple imputed datasets can be found in Appendix D.

Method 6 – Full information maximum likelihood without auxiliary variables: Makeup data were treated as missing. Full information maximum likelihood was employed using only the student pretest scores to aid in the estimation of parameters and

standard errors. The Mplus syntax employing FIML to estimate the parameters of interest (i.e., posttest mean, posttest variance, pretest-posttest covariance, and pre-post mean difference) can be found in Appendix F.

Method 7 – Full information maximum likelihood with university database and pretest auxiliary variables: Makeup posttest data were treated as missing. Full information maximum likelihood was employed using university database and pretest auxiliary variables. As mentioned previously, a typical assessment practitioner would not have access to posttest auxiliary variables if posttest makeup data were not collected. Thus, it was important to compare the results when including and excluding posttest auxiliary variables. Figure 3 provides a visual for this model and Appendix G provides the Mplus syntax.

Method 8 – Full information maximum likelihood with all auxiliary variables: Makeup posttest data were treated as missing. Full information maximum likelihood was employed, using all auxiliary variables. Figure 3 provides a visual for this model and Appendix H provides the Mplus syntax.

Auxiliary variables included in Methods 4, 5, 7, and 8 are displayed in Table 4. Posttest auxiliary variables are only included in Methods 5 and 8. When incorporating auxiliary variables (from either pretest or posttest) with missingness into the MI imputation model, auxiliary missing values were imputed along with posttest scores. Auxiliary variables were incorporated into FIML analyses utilizing a saturated correlates model (see Figure 3), which can handle auxiliary variables with missing values (Enders, 2008).

Comparing results. The results of these eight methods of analyzing the data were compared. The effectiveness of these different data analytic techniques should be dependent on the mechanism underlying the "missing" makeup data. If the data were determined to be MCAR, all eight methods should produce similar estimates of posttest mean, posttest variance, pretest-posttest covariance, and pre-post mean difference. However, the standard errors associated with these estimates should be slightly inflated.

If the data were determined to be MAR, we would expect the missing data methods designed to effectively handle MAR data (Methods 3-8) to be more similar to the complete dataset results (Method 1) than methods not designed for MAR data (Method 2). Further, methods including auxiliary variables (Methods 4, 5, 7, and 8) should provide greater accuracy (i.e., parameters and standard errors closer to those obtained from the complete data) than methods excluding auxiliary variables (Methods 3 and 6). As mentioned previously, the extent to which the inclusion of auxiliary variables reduces bias is dependent on the relationships between the included auxiliary variables, missingness, and posttest scores (Collins et al., 2001), which is examined in Research Question 2.

If the missingness mechanism were found to be MNAR, we should expect all methods of handling the missingness to differ from the complete dataset results. However, methods including auxiliary variables (Methods 4, 6, 7, and 8) that *partially* moderate the relationship between missingness and missing data values should affect parameter estimates and standard errors in the ways summarized in Table 2. Further, MI and FIML analyses excluding auxiliary variables (Methods 3 and 6) utilize the pretest scores of students with missing posttest scores in the estimation of the various parameter

estimates. Thus, even in a MNAR data situation, methods including auxiliary variables (Methods 4, 5, 7, and 8) should provide greater accuracy (i.e., parameters and standard errors closer to those obtained from the complete dataset) than methods excluding auxiliary variables (Methods 3 and 6), and all MAR-based methods (Methods 3-8) should provide greater accuracy than listwise deletion (Method 2).

Research question 4: Percent of missingness. Do the results associated with the previous research questions depend on the percent of missingness? If the eight approaches to handling missing data yield similar parameter estimates and standard errors, this result could be due to the low percentage of missingness associated with both datasets (5.9% for noncognitive test data and 5.5% for cognitive test data). To investigate this possibility, the analyses described above were repeated after the proportion of missingness was artificially inflated. This process was accomplished by randomly deleting student data from the Assessment Day sample to create datasets where missingness accounts for 25% or 50% of the complete data. This deletion was done while holding the makeup student data constant, so that makeup data accounted for 25% or 50% of the overall dataset. Thus, the missing data mechanism was held constant as the proportion of "missing" (i.e., makeup) data was increased.

For the noncognitive test sample, instead of the percentage of students who skipped the posttest equaling the observed 5.9% of the complete data, the percentage of students who attended a makeup session was 25% or 50% by reducing the proportion of students who initially attended the posttest. Thus, the "MAP 25% missingness" dataset consisted of 402 randomly selected Assessment Day attendees and the original 134 makeup attendees, for a total of 536 examinees (134/536 = 25% missing). The "MAP

50% missingness" dataset consisted of 134 randomly selected Assessment Day attendees and the original 134 makeup attendees, for a total of 268 examinees (134/268 = 50% missing). The NW-9 25% and 50% missingness datasets were constructed in a similar manner. Although this approach increases the missing data percentage while maintaining the missing data mechanism, reducing the number of Assessment Day attendees in the dataset also results in a reduction of overall sample size. Thus, the results should be interpreted cautiously. Previous simulation studies have commonly used 25% and 50% missingness (e.g., Collins et al., 2001). Importantly, missingness as high as 50% has occurred in educational testing programs such as NAEP (Chromy, 2005). Thus, these missingness percentages are realistic to many testing contexts.

The results of these analyses should help inform assessment practitioners that may have higher proportions of missing data. That is, practically small biases or standard error inflation at low missingness proportions may become problematically large at high missingness proportions. Thus, assessment practitioners encountering a high proportion of missingness due to nonattendance may need to adopt different approaches from assessment practitioners encountering lower missingness proportions.

Research question 5: Noncognitive versus cognitive. Do the answers to the previous research questions depend on whether the construct being examined is noncognitive or cognitive in nature? Parameters and standard errors associated with cognitive exam scores may be more affected by treating these scores as missing than parameters and standard errors associated with noncognitive measures. To assess this possibility, all of the analyses were conducted twice: once when modeling noncognitive test data (MAP scores) and again when modeling cognitive data (NW-9 scores).

If the results differed depending on whether cognitive or noncognitive data were being analyzed, best practices for handling posttest nonattendance missingness would depend on the construct being examined. For example, pre-post mean difference parameter estimates may be unbiased when multiply imputing posttest MAP scores, but biased when multiply imputing posttest NW-9 scores. In this case, assessment practitioners would be able to utilize MI for missing noncognitive posttest data, but would need to conduct makeup testing sessions for missing cognitive posttest data.

CHAPTER FOUR

Results

Noncognitive Measure (MAP) Results

Research question 1: Examining posttest response validity. A multiple group analysis indicated that posttest MAP scores from the makeup sample may have increased random responding. Low effort and random responding should reduce the MAP pre-post slope or increase the posttest residual variance, resulting in diminished posttest score validity for the makeup sample compared to the Assessment Day sample. Table 5 presents the pretest and posttest means and variances, as well as the freely estimated intercepts, pre-post slopes, and posttest residual variances for each group. The posttest mean was smaller and posttest variance was larger for the makeup sample than the Assessment Day sample. As would be expected if low motivation manifested in increased random responding, the pre-post slope was smaller and the posttest residual variance was larger for the makeup sample.

Table 6 presents the fit information for constraining the posttest intercepts, pre-post slopes, posttest residual variances, or all three to be equivalent across groups. The fit of Models 1 and 2 are sufficient, indicating that the posttest intercepts and pre-post slopes are equivalent across groups. However, Models 3 and 4 are associated with poor relative fit indices and statistically significant χ^2 tests, indicating the posttest residual variance is different across groups. The increased residual variance indicates the makeup examinees may have engaged in more rapid and thoughtless responding due to low motivation to perform.

Given the increased residual variance in posttest MAP scores for the makeup examinees, makeup student responses may be a less valid representation of student mastery approach orientation than Assessment Day student responses. Thus, parameter estimates obtained when *excluding* makeup student responses may be a more valid than those obtained when *including* makeup student responses. Specifically, including makeup posttest data could *bias* estimates of posttest variance, given posttest variance was inflated in the makeup sample. As a result, discrepancies between the variance when analyzing the complete (i.e., including makeup) dataset and the variance when treating makeup data as missing may not reflect true "bias" by the missing data handling techniques, but instead reflect the "bias" resulting from *including* invalid makeup posttest responses. Note also that this increased posttest variance may be a function of a subset of makeup examinees responding randomly, rather than the entire sample. The potential for invalid posttest MAP responses by the makeup sample will be considered in conjunction with the findings of the following research questions.

Research question 2: Examining the missing data mechanism. Bivariate relationships were examined between posttest attendance (*R*), posttest MAP scores (*Y*), and other measured dataset variables (*X*) to determine whether the MCAR assumption was met. Table 7 presents the descriptive statistics for these variables, and Table 8 presents the bivariate linear relationships. Note that pretest MAP score was significantly related to posttest MAP score (r = .382), but was not the strongest bivariate predictor of posttest MAP scores (*Y*). Posttest MAV and WAV scores were more strongly related to posttest MAP scores (r = .480 and -.500, respectively) than pretest MAP score, indicating that these posttest variables may need to be included as auxiliary variables to minimize

standard error inflation. Posttest attendance (*R*) was found to have a small but significant positive linear relationship with both pretest (r = .049) and posttest MAP scores (r = .138). Additionally, posttest attendance (*R*) was found to be significantly related to a number of other dataset variables, including gender, SAT verbal scores, GPA, pretest MAV, PAV, openness, conscientiousness, and agreeableness scores, and posttest PAP, PAV, WAV, conscientiousness, and agreeableness scores. Thus, compared to students attending Assessment day, the typical "makeup examinee" is more likely to be male with higher SAT verbal scores, lower GPA, lower mastery and performance orientation towards learning, higher work avoidance, and lower conscientiousness and agreeableness.

The significant bivariate relationships between posttest attendance (*R*) and both posttest MAP scores (*Y*) and other dataset variables (*X*s) indicated the MCAR assumption was violated. Further, all of the dataset variables that were related to posttest attendance (*R*) were also related to posttest MAP scores (*Y*). Thus, including these dataset variables as auxiliary variables should reduce the discrepancy between parameters obtained utilizing the complete dataset and those obtained utilizing MI or FIML, to the extent that these variables can moderate the relationship between missingness (*R*) and posttest MAP scores (*Y*), thereby transforming the MNAR mechanism to MAR. Note that this may not be reducing "bias", as the results of Research Question 1 indicate that the makeup scores may be biased themselves to an extent. That is, students' "true" levels of MAP are unknown, and thus true bias is difficult to assess.

The magnitudes of the correlations between posttest attendance (*R*) and the auxiliary variables (*X*s) were low in magnitude, ranging from r = -.083 to r = .110. Collins and colleagues (2001) have recommended auxiliary variables be included if they

are bivariately related to missingness or missing values above $r = |.4|$. As a result, including these auxiliary variables may not greatly moderate the *R*-*Y* relationship, and thus not reduce parameter bias to a great extent.

Nonlinear relationships with attendance (*R*) were also examined by comparing score distributions on all examined variables across Assessment Day attendees and makeup students. If a convex relationship was found between *R* and a dataset variable (i.e., missingness rates were higher at the high and low ends of the variable distribution), the dataset variable should be included as an auxiliary variable to reduce variance and covariance estimate bias. These density distributions are presented in Appendix I. No substantial nonlinear relationships were found between attendance (*R*) and any other examined variable.

Given the MCAR assumption was violated and missingness (*R*) was related to posttest MAP scores (*Y*), the partial linear correlation between posttest attendance (*R*) and posttest MAP scores (*Y*) was computed after controlling for different *individual* dataset variables (Table 9) and *sets* of dataset variables (Table 10) to assess the extent to which the MAR assumption was met. Examining Table 9, note the individual dataset variables that most moderated the relationship between posttest attendance (*R*) and posttest MAP scores (*Y*), resulting in a lower partial *R*-*Y* correlation, were all posttest variables (posttest WAV, Conscientiousness, and Agreeableness). Additionally, examining Table 10, the partial correlation between posttest attendance (*R*) and posttest MAP scores (*Y*) decreased as more dataset variables were added, and was lowest when posttest auxiliary variables were included. The reduced *R*-*Y* partial correlation when posttest auxiliary variables were included indicates that posttest auxiliary variables may need to be included to minimize

parameter bias. Additionally, the variance explained by the model including posttest auxiliary variables (R^2 = .526) was substantially higher than the variance explained by the model excluding posttest auxiliary variables (R^2 = .198), indicating that posttest auxiliary variables may need to be included to minimize standard error inflation. However, the partial correlation between posttest attendance (*R*) and posttest MAP scores (*Y*) remained significant after controlling for all dataset variables (partial *r* = .108), indicating that the MAR assumption was violated and the missingness mechanism can be considered MNAR. Additionally, this partial correlation (*r* = .108) was similar to the bivariate relationship between posttest attendance (*R*) and posttest MAP scores (*Y*) (*r* = .138), indicating that the auxiliary variables do not *greatly* moderate the relationship between *R* and *Y*. This small reduction in the partial correlation is not greatly surprising, given the weak relationships between the majority of auxiliary variables, posttest attendance (*R*), and posttest MAP scores (*Y*). Thus, the inclusion of these auxiliary variables is not likely to result in a substantial decrease in parameter "bias" (i.e., discrepancy between the complete dataset parameters and those obtained via MI or FIML procedures).

In addition to identifying the MNAR missing data mechanism, it was important to fully understand the models being used to account for the missing posttest values (i.e., makeup data) in the MI and FIML analyses. To this end, regression coefficients and squared semipartial correlations are presented for each of the auxiliary regression models. The two models examined include university database and pretest auxiliary variables excluding posttest auxiliary variables (Table 11) and including all potential auxiliary variables (Table 12). Examining these tables also provides an indication of the utility of each auxiliary variable (*X*) for predicting posttest MAP scores (*Y*) *after controlling for* all

other auxiliary variables. Comparing these results to the bivariate results (Table 8) and the partial correlations after controlling for each *individual* auxiliary variable (Table 9) presents a complicating and somewhat confusing picture of which auxiliary variables are "most important". For instance, pretest MAV score is significantly bivariately related to posttest MAP scores (r = .049), has a near-zero relationship with posttest MAP scores when other pretest auxiliary variables are included in the model (b = .000), which then becomes a significant *negative* slope when posttest auxiliary variables are included (b = -.038). This set of values showcases that, when the auxiliary variables (Xs) are placed in a model together, a combination of *moderator effects* (leading to a reduction in some predictor slopes) and *suppressor effects* (leading to an increase in some predictor slopes) complicates the interpretation of the relationships between the auxiliary variables (Xs) and posttest MAP scores (Y). Importantly, the simple bivariate relationships may not provide the best indication of which auxiliary variables should be included in the MI and FIML analyses.

Research question 3: Comparing missing data handling techniques.

Comparisons of parameters and standard errors obtained utilizing the complete dataset versus the missing data handling methods are presented in Table 13. No parameters obtained via any of the missing data handling techniques were substantially discrepant from the complete dataset parameters. Standardized discrepancy estimates ranged from -1.791 to 1.662 for listwise deletion. The utilization of MI or FIML (-1.742 to 1.431) and these techniques with auxiliary variables (-1.700 to 1.323) slightly reduced parameter discrepancy. Thus, the recommended inclusive analysis strategy (i.e., MI or FIML with all auxiliary variables) resulted in the lowest parameter discrepancy.

Standard error inflation was also minimal across all methods and parameters, with relative efficiency estimates ranging from 0.938 to 1.031 across all methods. MI and FIML did not offer substantial improvement in standard error inflation over listwise deletion and the inclusion of auxiliary variables with these techniques very slightly reduced standard error inflation for the majority of parameters. The minimal standard error inflation and bias may be due to the low percentage of missingness (5.9%). Thus, the parameters were estimated utilizing the various techniques with higher percentages of missingness.

Research question 4: Percentage of missingness. The 25% and 50% missingness datasets were obtained to determine the extent to which parameter bias and standard error inflation occurred at higher percentages of missingness. Parameters and standard errors obtained utilizing the 25% and 50% missingness datasets are presented in Tables 14 and 15. Standardized discrepancy estimates in the 25% missingness condition were large across all missing data handling techniques for all parameters except pretest-posttest covariance estimates. Across all missing data handling techniques, posttest mean and pre-post mean difference estimates were larger than the complete dataset, and posttest variance estimates were smaller than the complete dataset. As mentioned previously, increased random responding by makeup examinees manifested in a greater posttest residual variance for the makeup sample than the Assessment Day sample when predicting posttest scores from pretest scores. Thus, when posttest makeup data were treated as missing, posttest variance estimates were underestimated by the missing data handling techniques. Importantly, the positive relationship between posttest attendance and posttest MAP scores (i.e., MNAR) resulted in the overestimation of posttest mean

and pre-post mean difference parameters when makeup data were treated as missing. Notice that the utilization of MI or FIML techniques slightly improved posttest mean and variance estimates over listwise deletion. Moreover, the utilization of all auxiliary variables with MI or FIML reduced discrepancy estimates for all parameters except posttest variance estimates, which aligns with the reduced MNAR violations displayed in Table 10 (i.e., reduced *R-Y* partial correlation) when all auxiliary variables are included. Overall, the results suggest that the MCAR and MAR violations created substantial parameter discrepancies for the majority of parameters examined, which were ameliorated by utilizing advanced techniques (MI and FIML) with additional auxiliary variables. Additionally, standard error inflation was low across all parameters and handling techniques, and was lowest when all auxiliary variables were utilized in conjunction with MI or FIML.

The missing data handling techniques were more problematic in the 50% missingness condition. All parameters with the exception of pretest-posttest covariance estimates showed significant discrepancy from the complete dataset parameters utilizing all missing data handling techniques except MI with all auxiliary variables. Again, across all methods, posttest mean and pre-post mean difference estimates were larger than the complete dataset, and posttest variance estimates were smaller. The addition of auxiliary variables helped reduce these discrepancies for both MI and FIML techniques, as would be expected given the reduction in MNAR effects when auxiliary variables were included (Table 10). Additionally, standard error inflation was problematic for the majority of parameters and handling techniques. Overall, it appears that the extent of MCAR and MAR violations created significant issues for all missing data handling techniques in the

50% missingness condition, but these issues were somewhat ameliorated with the utilization of advanced techniques (MI and FIML) and additional auxiliary variables.

MAP results summary. Overall, the results from the MAP analyses conform to expectations given previous missing data research. Examining the partial correlations reveals that the addition of auxiliary variables (*X*s) *reduced* the partial correlation between posttest attendance (*R*) and posttest MAP scores (*Y*), but only slightly. The partial correlation remained significant after controlling for all auxiliary variables, indicating a MNAR mechanism. Posttest auxiliary variables accounted for a large proportion of variance in posttest MAP scores independent of other auxiliary variables (R^2 change = .328), with posttest MAV and WAV scores being strong bivariate predictors of posttest MAP scores. Accordingly, advanced missing data handling methods (MI and FIML) provided more accurate results than listwise deletion, and pursuing an *inclusive analysis strategy* (i.e., including more auxiliary variables) resulted in further accuracy. However, given the weak relationships between many auxiliary variables and missingness (*R*) and posttest scores (*Y*), including auxiliary variables did not greatly improve parameter estimates or standard errors overall. Given the MNAR mechanism, all techniques remained problematic at high proportions of missingness, with high parameter discrepancies and standard error inflation.

Cognitive Test (NW-9) Results

Research question 1: Examining posttest response validity. Similar to the MAP results, a multiple group analysis indicated that posttest NW-9 scores from makeup attendees may be compromised by decreased test-taking effort. This could manifest in a diminished NW-9 posttest intercept, a diminished pre-post slope, an increased posttest

residual variance for the makeup sample if low effort is resulting in diminished posttest score validity for the makeup sample compared to the Assessment Day sample. Table 16 presents the pretest and posttest means and variances, as well as the freely estimated intercepts, slopes, and residual variances for each group. The makeup sample has a lower intercept and a higher posttest variance, pretest-posttest slope, and posttest residual variance compared to the Assessment Day sample. Table 17 presents the fit information for constraining the posttest intercepts, pre-post slopes, posttest residual variances, or all three to be equivalent across groups. The model constraining all three parameters to be equivalent across groups was associated with a statistically and practically significant decline in fit, with the largest residuals associated with the posttest mean, indicating that the intercepts are not equivalent across groups. Thus, Assessment Day and makeup students differ in posttest NW-9 scores after controlling for their pretest NW-9 scores, with makeup students scoring lower at posttest. This difference may be due to makeup students responding randomly to items due to lower motivation, resulting in more incorrect answers.

Recall there was a greater residual variance associated with predicting posttest MAP scores from pretest MAP scores for makeup students compared to Assessment Day students. By contrast, makeup students had a lower predicted NW-9 posttest mean than Assessment Day students after controlling for pretest NW-9 score. When responding to MAP items, students rated their level of agreement with statements. Thus, random responding to posttest MAP items by makeup students would result in more variance in the ratings of agreement, resulting in an increased residual variance. NW-9 items are scored as correct or incorrect. In this instance, random responding to posttest NW-9 items

by makeup students would result in more incorrect items, leading to a lower NW-9 total score than would be predicted from their pretest score. Thus, both the NW-9 and MAP results suggest that makeup students responded more randomly or thoughtlessly at posttest than Assessment Day students.

Examining the density distributions of posttest NW-9 scores across groups (first graph in Appendix J) reveals that only a subset of makeup examinees may be responding randomly. That is, makeup posttest scores generally follow a negative skew, with only a few individuals scoring in the lower tail of the distribution. Thus, random responding may not be endemic to the entire makeup sample, and only a subset of makeup examinees are not putting forth effort on the NW-9 test.

Given the reduced posttest mean after controlling for pretest score for the makeup examinees, makeup student responses may be a less valid representation of student scientific reasoning knowledge than Assessment Day student responses. Thus, parameter estimates obtained when *excluding* makeup student responses may be a more valid representation of average student scientific reasoning knowledge and growth than those obtained when *including* makeup student responses. Specifically, posttest mean and pre-post mean change estimates may be biased by random responding in the makeup sample, given the lower intercept for that group compared to the Assessment Day attendee sample. As a result, discrepancies found between the parameters found when analyzing the complete (i.e., including makeup) dataset and the parameters found when treating makeup data as missing may not reflect true "bias" by the missing data handling techniques, but instead reflect the "bias" resulting from *including* invalid makeup posttest

responses. The potential for invalid posttest NW-9 responses by the makeup sample will be considered in conjunction with the findings of the following research questions.

Research question 2: Examining the missing data mechanism. Bivariate relationships were examined between posttest attendance (*R*), posttest scientific reasoning scores (*Y*), and other measured dataset variables (*X*) to determine whether the MCAR assumption was met. Table 18 presents the descriptive statistics for these variables, and Table 19 presents the bivariate linear relationships. As expected, pretest NW-9 scores had the strongest bivariate relationship with posttest NW-9 scores (r = .663). Given the magnitude of this relationship, it is possible that the auxiliary variables (*X*s) may not account for additional independent variance in posttest NW-9 scores (*Y*) after controlling for pretest NW-9 scores, and thus may not be important to gather.

Although posttest attendance (*R*) had a nonsignificant negligible linear relationship with pretest (r = -.043) and posttest scientific reasoning scores (r = .059), posttest attendance (*R*) was significantly linearly related to gender, SAT verbal scores, pretest and posttest MAP scores, pretest Conscientiousness and MAI-R scores, and posttest WAV scores. The significant bivariate relationships between posttest attendance (*R*) and other dataset variables (*X*s) indicated that the MCAR assumption was violated. Further, gender and SAT verbal scores were also significantly linearly related to posttest scientific reasoning scores (*Y*). Thus, including gender and SAT verbal as auxiliary variables should reduce parameter bias (given each variable was related to *both* missingness and scientific reasoning scores). However, although the magnitude of the relationship with posttest scientific reasoning scores was non-negligible (gender r = .169; SAT Verbal r = .536), the magnitudes of the correlations were low between posttest

attendance (*R*) and gender (r = -.105) and SAT Verbal scores (r = -.081). Thus, although the MCAR assumption is violated in a statistical sense, relatively little parameter bias may result from excluding these auxiliary variables from MI and FIML analyses.

Nonlinear relationships were also examined by comparing score distributions on all examined variables across Assessment Day attendees and makeup students. If a convex relationship was found between *R* and a dataset variable (i.e., missingness rates were higher at the high and low ends of the variable distribution), the dataset variable should be included as an auxiliary variable to reduce variance and covariance estimate bias. These density distributions are presented in Appendix J. No substantial nonlinear relationships were found between attendance (*R*) and any other examined variable.

Given that the MCAR assumption was violated, the partial linear correlation between posttest attendance (*R*) and posttest scientific reasoning scores (*Y*) was computed after controlling for different *individual* dataset variables (Table 20) and *sets* of dataset variables (Table 21) to assess the extent to which the MAR assumption was met. Interestingly, the partial correlations between posttest attendance (*R*) and posttest scientific reasoning scores (*Y*) were *greater* than the bivariate relationship between *R* and *Y* (r = .059) after controlling for some individual dataset variables (pretest NW-9 scores, gender, SAT scores; see Table 20) and sets of dataset variables (increasing to .117 after controlling for pretest NW-9 scores, and to .149 after controlling for both pretest NW-9 scores and university database variables; see Table 21). When the MAR assumption is typically discussed (e.g., Enders, 2010) or simulated (e.g., Collins et al., 2001), there is usually a significant bivariate relationship between missingness (*R*) and the variable with missing values (*Y*) that is *spurious* due to a shared relationship with another variable (*X*).

When this other variable (*X*) is controlled for, the partial relationship between missingness *R* and *Y* diminishes or disappears. However, in the current study, the partial relationship between *R* and *Y increases* as a result of controlling for other dataset variables. These findings indicate that *statistical suppression* is occurring when only the bivariate correlation is examined.

Suppression is an oft-discussed statistical phenomenon in social science research (e.g., MacKinnon, Krull, & Lockwood, 2000) that can be difficult to understand. A suppressor variable is defined as

> a variable which increased the predictive validity of another variable (or set of variables) by its inclusion in a regression equation… Thus, a suppressor variable is not defined by its own regression weight but rather by its effects on other variables in a regression system. (Conger, 1974, pp. 36-37)

For example, when pretest scientific reasoning score was added to the model, the partial correlation between posttest attendance (*R*) and posttest scientific reasoning scores (*Y*) was larger (.117) than the bivariate correlation between attendance and posttest scientific reasoning scores (.059). Thus, pretest scientific reasoning score was a *suppressor variable* for posttest attendance (*R*) in the prediction of posttest scientific reasoning scores (*Y*). This larger partial correlation is due to the pretest scientific reasoning scores having a negative relationship with posttest attendance (i.e., those with higher pretest scores are less likely to attend Assessment Day), but a positive relationship with posttest scientific reasoning scores (*Y*) (i.e., those with higher posttest scores are more likely to attend Assessment Day), as is evident when examining the partial correlation formula:

$$r_{RY.X} = \frac{r_{RY} - r_{RX} r_{YX}}{\sqrt{1 - r_{RX}^2}\sqrt{1 - r_{YX}^2}} \quad (3)$$

Inserting the correlations between posttest attendance (*R*), posttest scientific reasoning (*Y*) and pretest scientific reasoning (*X*) from Table 8 gives:

$$r_{RY.X} = \frac{.059 - (-.043) * (.663)}{\sqrt{1-(-.043)^2}\sqrt{1-(.663)^2}} \quad (4)$$

$$r_{RY.X} = \frac{.059 - (-.02851)}{(.99908) * (.74862)} \quad (5)$$

$$r_{RY.X} = .118 \quad (6)$$

Conceptually, the bivariate relationship between posttest attendance (*R*) and posttest scientific reasoning scores (*Y*) *ignores* pretest scientific reasoning scores (*X*). That is, if the Assessment Day and makeup samples mean posttest scientific reasoning scores (*Y*) were compared there would not be a significant difference between the mean scores of the two groups. However, when pretest scientific reasoning score (*X*) is entered into the regression equation, the partial correlation quantifies the relationship between posttest attendance (*R*) and posttest scientific reasoning scores (*Y*) with pretest scientific reasoning score (*X*) *held constant* (Edwards, 1976). Thus, *at each level of* pretest scientific reasoning score (*X*), there is a significant positive relationship between posttest attendance (*R*) and scientific reasoning score (*Y*) - given equivalent pretest scores, students attending Assessment Day at posttest are significantly higher on posttest scientific reasoning than students attending makeup.

Gender serves as an example of a categorical suppressor variable. Gender is negatively related to posttest attendance (r = -.105), but positively related to posttest scientific reasoning scores (r = .169). That is, men are less likely than women to attend their assigned assessment session at posttest, but score higher on average on the NW-9 test than women. Thus, when the bivariate relationship between posttest attendance (*R*)

and posttest scientific reasoning scores (*Y*) is examined, gender (*X*) is *ignored* and there appears to be no relationship. However, *at each level of gender* (i.e., examining only males and examining only females), posttest attendance (*R*) and posttest scientific reasoning scores (*Y*) have a significant positive relationship. That is, women attending Assessment Day score higher than women attending makeup testing, and men attending Assessment Day score higher than men attending makeup testing. As a result, when gender (*X*) is included in the regression model, the partial correlation between posttest attendance (*R*) and posttest scientific reasoning scores (*Y*) increases.

In sum, the partial correlations between posttest attendance (*R*) and posttest scientific reasoning scores (*Y*) indicated the MAR assumption was violated. Interestingly, the extent to which the mechanism could be considered MNAR (i.e., missingness related to Y) actually *increased* as more auxiliary variables were included in the regression model due to a number of suppressor variables present in the model (e.g., gender). This pattern mirrors those described in previous missing data simulation research (Thoemmes & Rose, in press) where conditioning on some auxiliary variables led to an increased *R*-*Y* covariance. In this previous research, inclusion of these auxiliary variables in MI or FIML analyses led to biased mean estimates. Thus, Thoemmes and Rose (in press) labeled these *bias-inducing* variables. Thoemmes and Rose (in press) also identified a number of alternative configurations where an auxiliary variable may introduce dependencies between *R*, *Y*, and unobserved variables related to *R* or *Y* themselves. Thus, suppression effects are only one kind of configuration that can result in biasing effects. Given the finding of a suppression mechanism, mean estimates may be biased when including these bias-inducing auxiliary variables (e.g., gender).

In addition to identifying the MNAR missing data mechanism, it was important to fully understand the models being used to account for the missing posttest values (i.e., makeup data) in the MI and FIML analyses. To this end, regression coefficients and squared semipartial correlations are presented for each of the auxiliary regression models. The two models examined include university database and pretest auxiliary variables excluding posttest auxiliary variables (Table 22) and including all potential auxiliary variables (Table 23). Examining these tables also provides an indication of the utility of each auxiliary variable (*X*) for predicting posttest NW-9 scores (*Y*) *after controlling for* all other auxiliary variables. Similarly to the MAP results, comparing the results of the regression models presented in Tables 22 and 23 to the bivariate relationships (Table 19) and the partial correlations after controlling for each *individual* auxiliary variable (Table 20) presents an unclear picture of which auxiliary variables are "most important". For instance, pretest WAV score is not significantly bivariately related to posttest NW-9 scores (r = -.029), and is not a significant predictor of posttest NW-9 scores when only pretest auxiliary variables are included (b = .095), but it becomes a significant *positive* predictor when posttest auxiliary variables are included (b = .105). As with the MAP results, when the auxiliary variables (*X*s) are placed in a model together, a combination of *moderator effects* (leading to a reduction in some predictor slopes) and *suppressor effects* (leading to an increase in some predictor slopes) complicates the bivariate relationships between the auxiliary variables (*X*s) and posttest NW-9 scores (*Y*), and the simple bivariate relationships may not provide the best indication of which auxiliary variables should be included in the MI and FIML analyses. Additionally, it is unclear how the suppression effects that lead to an *increased R-Y* partial correlation after controlling for

the different auxiliary sets will affect parameter bias and standard errors when these auxiliary variables are included in MI and FIML analyses.

Research question 3: Comparing missing data handling techniques.

Comparisons of parameters and standard errors obtained utilizing the complete dataset versus the missing data handling methods are presented in Table 24. As would be expected given both the low rate of missingness (5.5%) and the weak relationships between posttest attendance (*R*) and *both* posttest scores (*Y*) and other variables (*X*), no parameters obtained via any of the missing data handling techniques were substantially discrepant from the complete dataset parameters. Standardized discrepancy estimates ranged from -0.685 to .888. Note, however, that the addition of university database and pretest auxiliary variables slightly increased standardized discrepancy estimates for both the posttest mean and pre-post mean difference estimates when utilizing MI or FIML. This slight increase in discrepancy is likely due to the *R*-*Y* dependencies introduced by certain variables noted above (e.g., gender), given previous research has found similar effects (Thoemmes & Rose, in press).

Standard error inflation was also minimal, with relative efficiency estimates ranging from 0.964 to 1.028. Utilizing advanced missing data handling methods (MI or FIML) slightly reduced standard error inflation compared to listwise deletion. However, the inclusion of auxiliary variables did not consistently reduce standard errors. This lack of standard error improvement may be due to pretest NW-9 score (which is included in the no-auxiliary MI and FIML models) being highly correlated with posttest NW-9 scores (r = .663). Thus, the inclusion of auxiliary variables resulted in a comparatively small improvement in the prediction of posttest NW-9 scores (model R^2 improving from .440 to

.562). This finding is in contrast to the MAP results, where auxiliary variable inclusion resulted in a large increase in the proportion of variance explained in posttest MAP scores. As a result, the inclusion of auxiliary variables had little effect on standard error inflation associated with the NW-9 parameters. Additionally, the minimal standard error inflation and bias may be due to the low percentage of missingness (5.5%). Thus, it was important to proceed with estimating parameters and standard errors at higher rates of missingness.

Research question 4: Percentage of missingness. The 25% and 50% missingness datasets were created to determine the extent to which parameter bias and standard error inflation occurred at higher percentages of missingness. Parameters and standard errors obtained utilizing the 25% and 50% missingness datasets are presented in Tables 25 and 26. As expected, standardized discrepancy estimates in the 25% missingness condition were larger than in the 5.5% missingness condition, but were not large in an absolute sense. Again, discrepancy estimates were slightly higher for posttest mean and pre-post mean difference estimates when university database and pretest auxiliary variables were included in MI and FIML analyses. Standard error inflation was minimal for most parameters and handling techniques. However, standard error inflation was problematic for pre-post mean difference estimates across conditions. Inflation was lower than other handling methods for MI utilizing university database and pretest auxiliary variables, but this result may have been idiosyncratic of the 20 imputations used. In the other conditions, pre-post mean difference relative efficiency estimates ranged from 1.211 to 1.266. Thus, the sample size would have to be increased by

between 21.1% and 26.6% to obtain the same standard errors utilizing the missing data techniques that were obtained utilizing the complete dataset.

As expected, standardized discrepancy and relative efficiency estimates were larger in the 50% missingness condition compared to the 5.5% and 25% conditions. Additionally, the FIML analysis including all auxiliary variables did not converge. As mentioned previously, FIML analyses with large numbers of auxiliary variables can create estimation problems (Savalei & Bentler, 2009). Thus, the nonconvergence in the all-auxiliary FIML analysis may be due to the large number of auxiliary variables relative to the number of individuals in this sample (20 auxiliary variables and 92 cases). Posttest mean estimates were greatly positively discrepant (i.e., estimates were larger than those obtained analyzing the complete dataset) utilizing all methods except listwise deletion. Additionally, posttest mean standard error inflation was large when utilizing listwise deletion. Posttest variance estimates were greatly negatively discrepant (i.e., estimates were smaller than those obtained analyzing the complete dataset) utilizing all methods except MI with auxiliary variables, and standard error inflation was large when utilizing MI with all auxiliary variables. Pre-post mean change was greatly positively discrepant (i.e., estimates were larger than those obtained analyzing the complete dataset) when utilizing all missing data handling techniques, and standard errors were substantially inflated. As mentioned previously, increased random responding by makeup examinees manifested in lower posttest NW-9 scores than would be expected given their pretest NW-9 scores. Thus, when makeup posttest data are treated as missing, posttest mean and pre-post mean difference parameters are overestimated. Overall, it appears that excluding

makeup students results in smaller variance and covariance estimates and larger posttest mean and pre-post mean difference estimates.

Utilizing advanced techniques with additional auxiliary variables appears to provide more accurate variance and covariance estimates, but *less* accurate posttest mean and pre-post mean difference estimates. The decreased mean and mean difference accuracy is most severe when only university database and pretest auxiliary variables are included in conjunction with MI or FIML. Note that the condition only utilizing university database and pretest auxiliary variables also resulted in one of the largest partial correlations between posttest attendance and posttest scientific reasoning scores (partial = .143; see Table 21). Thus, including bias-inducing suppressor auxiliary variables (e.g., gender) that lead to an *increased* partial correlation between posttest attendance (*R*) and posttest NW-9 scores (*Y*) appear to have resulted in increased posttest mean and pre-post mean difference discrepancies.

NW-9 results summary. The results of the NW-9 analyses reinforce important issues regarding the treatment of missing data when encountering induced dependencies between missingness (*R*) and missing values (*Y*) when including some auxiliary variables. As noted by Thoemmes and Rose (in press), including auxiliary variables that introduce dependencies between missingness (*R*) and missing values (*Y*) can bias mean estimates. Posttest attendance (*R*) was found to be bivariately unrelated to posttest scientific scores (*Y*), but was found to have a larger *partial* correlation after controlling for auxiliary variables (*X*s). These partial correlations were still small in absolute magnitude (with the largest being .149), thus MAR violations were practically small. Given the small MAR violations and low percentage of missing data (5.5%), utilizing any missing data

treatment method (listwise, MI, or FIML) did not result in substantial parameter discrepancies or standard error inflation when compared to the complete dataset.

Additionally, multiple group analyses revealed that makeup student responses at posttest may not be valid, due to makeup students achieving lower scores at posttest than would be predicted given their pretest scores. Thus, the current method of dealing with makeup students (i.e., listwise deletion) may not be problematic, and may actually be beneficial. However, standard error inflation became problematic for pre-post mean difference estimates when missingness was increased to 25%, and parameter discrepancy and standard error inflation both became problematic when missingness was increased to 50%. Thus, even small MCAR or MAR violations can be problematic when combined with large missingness percentages.

Further, posttest mean and pre-post mean difference estimates were *more* discrepant when auxiliary variables were included in the analysis, suggesting that including auxiliary variables that introduce *R*-*Y* dependencies may increase bias when they are included in MI or FIML analyses. As a consequence of the findings regarding bias-inducing variables, following the *inclusive analysis strategy* that is currently recommended (Collins et al., 2001) may *not* be the best approach if the auxiliary variables included in the MI or FIML analyses are introducing dependencies in the *R*-*Y* relationship. Whereas the inclusive analysis strategy resulted in reduced parameter bias and standard error inflation in the MAP analyses (as expected), the inclusion of suppressor auxiliary variables in the NW-9 analyses led to an *increased* partial correlation between missingness (*R*) and posttest NW-9 scores (*Y*), and *increased* bias in parameter estimates. Unfortunately, in most applied missing data situations (i.e., where

the researcher does not have access to the missing data), the researcher will *not* know whether included auxiliary variables will introduce *R*-*Y* dependencies. As a result, the findings in this study relevant to suppressor variables cast some doubt on the inclusive analysis strategy.

CHAPTER FIVE

Discussion

The results of this study provide useful guidelines for assessment practitioners who face missing data issues due to nonattendance. Following the recommendations of Graham (2009), initially missing scores were recovered to determine the exact missing data mechanism and the bias introduced by various missing data handling techniques. The following results emerged, which are quickly summarized here and discussed below. First, there was evidence that makeup responses possessed questionable validity for both noncognitive and cognitive measures. This may have been true for only a subset of examinees. Second, the missing data mechanism underlying posttest nonattendance was found to be MNAR for both noncognitive and cognitive tests. For the noncognitive test, this MNAR mechanism resulted in predictable analysis results when comparing missing data handling techniques, as the inclusive analysis strategy (i.e., MI or FIML with auxiliary variables) yielded lower parameter "bias" (i.e., discrepancy from the complete dataset results) and reduced standard error inflation. Again, note that we do not know if this is true "bias", as we do not know true student MAP levels. Interestingly, for the cognitive test, a number of dataset variables (e.g., gender) introduced *R*-*Y* dependencies, in that partialling their effects out of both posttest nonattendance (*R*) and posttest scientific reasoning scores (*Y*) resulted in a *stronger R – Y* relationship. Posttest mean and pre-post mean difference estimates in the cognitive sample were more positively "biased" (i.e., more discrepant from the complete dataset estimates), although posttest variance and pre-post covariance estimates were improved. This reinforced recent research into bias-inducing auxiliary variables, where including some auxiliary variables in MI and

FIML analyses slightly *increased* bias in mean estimates (Thoemmes & Rose, in press). Additionally, utilizing MI or FIML techniques or including additional auxiliary variables did not consistently reduce standard error inflation for the cognitive test. This lack of improvement is not surprising given the weak relationships between the various auxiliary variables and posttest NW-9 scores. Third, although parameter "bias" (i.e., discrepancy from the complete dataset results) and standard error inflation were not problematic for either the noncognitive or cognitive test when makeup data were treated as missing, this finding appeared to be the result of low missingness percentages. When missingness percentages were artificially increased to 25% and 50%, significant parameter bias and standard error issues became apparent across missing data handling techniques.

Reduced Posttest Score Validity

Given the results of the multiple group analyses, there is some evidence that makeup posttest responses may have been affected by lower motivation and random responding. In the noncognitive sample, increased posttest score variance that was unrelated to pretest scores suggests that makeup examinees may have engaged in random or thoughtless responding at higher rates than the Assessment Day sample. In the cognitive sample, lower posttest scores for the makeup sample than would be predicted by their pretest scores suggests that random or thoughtless responding resulted in more incorrect answers. Examining the variable density distributions for posttest cognitive scores revealed that this reduced motivation may only be problematic for a subset of makeup examinees. Thus, assessment practitioners at the university under study should consider continuing to exclude makeup testing results from overall educational accountability estimates until the validity of makeup posttest responses can be further

studied, and if deemed problematic, improved. If future studies determine that makeup posttest responses are affected by careless or random responding, including makeup student data could be considered invalid, and including these data could *bias* estimates of pre-post growth.

Note that test-taking motivation was *measured* at posttest via the SOS measure. Thus, if the problematic multiple group results were the product of decreased motivation at posttest by the makeup sample, we would expect posttest attendance and posttest effort scores to be positively correlated. However, this was not true for either the MAP or NW-9 sample, as posttest effort scores were not significantly related to posttest attendance. Previous research has found that test-taking effort can vary substantially over the course of a testing period (Barry, 2010; Barry, Horst, Finney, Brown, & Kopp, 2010; Horst, 2010). However, test-taking effort was measured once at the end of the testing session. Thus, one overarching test-taking effort score may not be sensitive to the lack of motivation on any single measure. Measuring test-taking effort after each instrument may provide a more accurate representation of test-taking effort, and these test-specific effort scores may be useful as future auxiliary variables. Additionally, recent research (Finney, Sundre, Swain, & Williams, 2014) suggests that the *change* in effort scores from pretest to posttest is more predictive of scores than their absolute value. Thus, filtering on motivation *change* may result in more accurate value added scores.

If future work uncovers that makeup students are not providing valid responses, steps could be taken to improve test-taking motivation. Previous research has found proctoring to have an effect on student test-taking effort levels and test scores (Lau, Swerdzewski, Jones, Anderson, & Markle, 2009). Thus, modifying makeup testing

proctoring to enhance motivation may result in more valid responses. For instance, holding makeup testing sessions on Fridays and Saturdays may be leading to decreased makeup student motivation. An alternative testing time may be considered to obtain more valid responses.

MNAR Mechanism, Suppressor Effects, and Missing Data Handling

Although the missing data mechanism was found to be MNAR for both the noncognitive and cognitive tests, the nature of the MNAR mechanism was vastly different. For the noncognitive test, posttest attendance (*R*) was bivariately related to posttest MAP scores (*Y*) indicating an MNAR mechanism. However, this relationship was slightly moderated by the variables in the dataset, resulting in a decreased partial correlation between *R* and *Y* when dataset variables were included in the model. In particular, posttest MAV and WAV scores were strong bivariate predictors of posttest MAP scores, and thus were important to include as auxiliary variables in MI and FIML analyses. The addition of auxiliary variables decreased parameter bias and standard error inflation. Thus, the noncognitive test results appear to affirm the inclusive analysis strategy as the relationships between missingness, posttest scores and auxiliary variables aligned with the typical simulation work that assesses the utility of the inclusive strategy.

By contrast, the results of the cognitive test analyses appear to challenge the inclusive analysis strategy. Although posttest attendance (*R*) was not related to posttest NW-9 scores (*Y*) bivariately, the partial correlation between these two variables increased as additional dataset variables were partialled out of both variables. This increased partial correlation was due to some suppressor variables, such as gender, increasing the relationship between *R* and *Y* when there were included in the model. The presence of

induced *R*-*Y* dependencies made it difficult to determine the most important auxiliary variables to include in MI and FIML analyses. Given these suppressor auxiliary variables increase the relationship between missingness (*R*) and posttest scores (*Y*), it follows that including these auxiliary variables may result in increased parameter bias. Accordingly inclusion of these additional auxiliary variables decreased variance and covariance estimate bias, but increased mean and mean difference bias.

The findings associated with the cognitive test results confirm previous work examining bias-inducing auxiliary variables (Thoemmes & Rose, in press). Instances where the partial relationship between missingness (*R*) and the variable with missing values (*Y*) *increases* as additional auxiliary variables are included in the model, has only recently been explored. The results of this research and previous work by Thoemmes & Rose (in press) indicate that including suppressor auxiliary variables in an analysis *increases* the bias of some parameters (e.g., mean and mean difference estimates), while *decreasing* the bias of other parameters (e.g., variance and covariance estimates). The effects of these suppressor auxiliary variables on standard error estimates were unclear in the current study. Additionally, no research has examined the effects of suppressor auxiliary variables that increase the predictive utility of *other auxiliary variables* for the variable with missing values (*Y*).

Percentage of Missingness

Predictably, results became more problematic as missingness percentages increased. In the noncognitive sample, all parameters except pretest-posttest covariance became increasingly biased as the missingness percentage increased, and all standard errors became inflated. Bias and standard error inflation were partially ameliorated by

utilizing advanced techniques (MI and FIML) combined with auxiliary variables. However, as would be expected given the MNAR mechanism, the utilization of these techniques did not completely eliminate parameter bias. Although parameter estimates and standard error inflation became similarly problematic at higher missingness percentages for the cognitive sample, the utilization of MI or FIML with auxiliary variables only served to increase the bias of some parameters (posttest mean and pre-post mean difference).

From examining the 25% and 50% missingness results, it becomes apparent that any issues with missing data handling techniques become more exaggerated at higher percentages of missingness. Although not directly addressed in this study, it is also likely that the missing data mechanism will be *different* at higher percentages of missingness. That is, the causes of 25% or 50% missingness are likely different and more severe (i.e., more likely to be MNAR) than the causes of 5% or 6% missingness. For instance, high rates of twelfth grade NAEP survey dropout were found to be the product of a myriad of nonrandom sources, including private school nonparticipation and lack of student attendance or motivation in low-income and urban school districts (Chromy, 2005). Thus, it is imperative that studies such as the current one that examine the *rate*, *mechanism*, and *potential bias* of missingness be conducted to thoroughly understand any missingness that may occur in educational accountability contexts.

Limitations and Future Research Directions

This study had a number of strengths, including collecting previously missing data to empirically determine the exact missing data mechanism and the bias introduced by utilizing various missing data handling techniques. However, there are a number of

limitations to note. The missing data percentages were low (between 5% and 6%) in this study. Although datasets with higher missingness percentages were artificially constructed, it is unclear whether a real dataset with 25% or 50% missingness (e.g., some NAEP data; Chromy, 2005) would exhibit similar bias and standard error inflation patterns. Additionally, the datasets with higher missingness percentages were formed by randomly deleting Assessment Day attendee cases, resulting in a lower overall sample size. As a result, it is unclear whether some of the results in the 25% and 50% missingness conditions are a consequence of increased missingness percentages or a lower overall sample size. This study also examined missingness in one higher education assessment context in one university. Thus, assessment practitioners should not assume the mechanisms underlying the posttest nonattendance missingness in this study will extend to other missing data situations.

Although the results provide some indication that including auxiliary variables that induce *R*-*Y* dependencies may create problems for the inclusive analysis strategy, future research must be done in this area. Research has only recently focused on this issue (Thoemmes & Rose, in press). Thus, these results should be replicated in other situations where induced *R*-*Y* dependencies are suspected to underlie a missingness mechanism. Specifically, it would be useful to determine the effects of the dependencies on standard errors. Additionally, if future findings further challenge the inclusive analysis strategy, concrete recommendations regarding auxiliary variable inclusion should be determined based on results.

Future researchers are encouraged to also heed Graham's (2009) advice and conduct studies to determine the exact mechanism underlying the missingness in other

missing data situations. These studies will inform the best method to handle such missingness, and help ensure that results from education assessments are as accurate and informative as possible.

Implications for Policy Makers

In this study, missingness rates were low and did not introduce a large amount of bias in student growth estimates. However, even slight differences in value-added estimates can have large implications for educational policies. For instance, institutions, programs, and even individual teachers or faculty can be held accountable based on their value-added estimates. A slight difference due to missingness could result in a program's funding being cut or a faculty member being dismissed. Thus, policy makers should interpret value-added estimates in the presence of missingness carefully. The percentage of missingness, the likely underlying missing data mechanism, and the missing data treatment method used when analyzing the data should all be carefully considered when evaluating value-added estimates. These issues are outlined well by Chromy (2005), who recommends introducing incentives to limit missingness so that these missingness issues only occur to a small extent.

Implications and Recommendations for Assessment Practitioners

Assessment practitioners must acknowledge that missing data constitute a considerable problem for educational assessment and missing data issues do not have any "quick fixes." The assessment practitioner is advised, then, to *endeavor to limit missingness if possible*. As noted above, one possible reason for the lack of bias or standard error inflation is the low percentage of missingness (5-6%). At the university where this study was conducted, the percentage of students attending their assigned

Assessment Day testing session has increased dramatically over the years. Much of this is due to concerted efforts to communicate testing times and obligations to students via multiple pathways (e.g., email, campus advertisements). In addition, students have a concrete incentive to complete their assessments, as the university will place a hold on their academic record if they do not complete them. As noted by Chromy (2005), having firm and clear contingencies related to test completion can dramatically increase response rates. Thus, the assessment practitioner may be best advised to fix missingness (by limiting or eliminating it) on the front-end, rather than trying to compensate for large amounts of missingness after assessments have been administered and data have been collected.

If missing data is unavoidable, through reporting of missing data and its extent is a *minimum* standard that assessment practitioners should adopt. Failure to report or acknowledge missingness is an ethical issue, as results could be misinterpreted (Enders & Gottschall, 2011). Responsible missingness documentation involves reporting both the *extent* and the possible *causes* of missingness. Reporting the extent and cause of the missingness allows assessment results to be interpreted within the context of the missing data situation. As noted by Enders and Gottschall (2011), reporting the cause of the missingness may limit missingness or improve missing data handling in future research. For instance, if some personality or developmental traits are suspected to increase the likelihood of posttest nonattendance (e.g., entitlement, reactance), those variables could be collected at pretest to serve as auxiliary variables. In this manner, adhering to more rigorous reporting requirements related to missing data could lead to improvements in assessment design.

Often, the potential causes of missingness may not be immediately clear. In these cases, assessment practitioners should *attempt to understand the missingness that exists by collecting plentiful data***.** In this study, the mechanism underlying missingness was uncovered by examining the relationships between missingness (*R*), the variable with missing values (*Y*), and additional dataset variables (*X*s). By collecting this information, a "profile" was established of the typical makeup examinee. This profile can then be used to design interventions to prevent nonattendance in the future. For instance, students missing at posttest were found to be lower on academic motivation and conscientiousness, while higher in work avoidance. Thus, the makeup student profile is one of a generally unmotivated student. Given this profile, communications with students to encourage Assessment Day attendance may target motivation directly, possibly by appealing to students' sense of academic citizenship, or their responsibility to the university (Wise, 2009).

In the current study, understanding the variables that related to missingness (*R*) and the missing values themselves (*Y*) also allowed for useful hypotheses regarding how different missing data handling techniques may be biased. As a part of collecting plentiful data, assessment practitioners should *recover some or all of the missing data for one or several cohorts, to empirically determine the missing data mechanism in their specific testing context*. Again, this will help identify the best way to handle the missingness in that particular setting, potentially help minimize the missingness in the future, and help to inform missing data research.

Overall, the results also indicate that the applied assessment practitioner *should not make assumptions regarding the absolute best way to handle missingness.* Although

the inclusive analysis strategy is generally advisable, the results of this study indicate that one analysis strategy may not fit all missing data situations. However, pending further research, *it is still advisable to utilize MI or FIML with auxiliary variables over listwise deletion in the majority of missing data situations.*

Assessment practitioners may be able to overcome substantial missing data issues by following the five strategies listed above: 1) attempt to limit missingness, 2) thoroughly document missingness rates and causes when it occurs, 3) attempt to understand missingness by collecting plentiful data, 4) further attempt to understand missingness by recovering some or all initially missing data, 5) generally utilize MI or FIML with auxiliary variables, but be cautious not to assume that missingness can be adequately handled in all data situations with this inclusive analysis strategy. Overall, more research is needed on the missing data handling techniques examined in this study, as well as on more novel techniques (e.g., MNAR-based techniques), to provide increasing accuracy in missing data situations. However, the recommendations above provide useful guidance for assessment practitioners given the current state of missing data research.

References

Allison, P. D. (2002). *Missing data.* Thousand Oaks, CA: Sage.

Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational Researcher, 37,* 65-75.

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumaker (Eds.), *Advanced structural equation modeling* (pp. 243-277). Mahwah, NJ: Lawrence Erlbaum.

Bandalos, D. L., & Leite, W. (2013). Use of Monte Carlo studies in structural equation modeling research. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 625-666). Charlotte, NC: Information Age.

Baraldi, A. N., & Enders, C. K. (2012). Missing data methods. In T. Little (Ed.), *Oxford handbook of quantitative methods: Vol. 2.* New York: Oxford University Press.

Barry, C. L. (2010). *Examining change in motivation across the course of a low-stakes testing session: An application of latent growth modeling* (Doctoral dissertation, James Madison University).

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing, 10,* 342-363.

Beale, E. M. L., & Little, R. J. A. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society, Series B, 37,* 129-145.

Brown, R. L. (1994). Efficacy of the indirect approach for estimating structural equation models with missing data: A comparison of five methods. *Structural Equation Modeling: A Multidisciplinary Journal, 1,* 287-316.

Brown, A. R. & Finney, S. J. (2011). Low-stakes testing and psychological reactance:

Using the Hong Psychological Reactance Scale to better understand compliant and non-compliant examinees. *International Journal of Testing, 11*, 248 - 270.

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society, Series B, 22,* 302-306.

Chromy, J. R. (2005). *Participation standards for 12th grade NAEP*. Washington, D. C.: National Assessment Governing Board. Retrieved May 17, 2013, from www.nagb.org/publications/chromy_paper_revised.doc

College Board (2012). *Trends in college pricing, 2012.* Retrieved from http://trends.collegeboard.org/sites/default/files/college-pricing-2012-full-report-121203.pdf

Collins, L. M., Schafer, J. L., & Kam, C-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6,* 330-351.

Conger, A. J. (1974). A revised definition for suppressor variables: A guide to their identification and interpretation. *Educational and Psychological Measurement, 34*, 35-46.

Crawford, S. L., Johnson, W. G., & Laird, N. M. (1993). Bayes analysis of model-based methods for nonignorable nonresponse in the Harvard Medical Practice Survey. *Case Studies in Bayesian Statistics, 83,* 78-117.

Demirtas, H., & Schafer, J. L. (2003). On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine, 22,* 2553-2575.

Edwards, A. L. (1976). *An introduction to linear regression and correlation.* San

Francisco: W. H. Freeman & Co.

Enders, C. K. (2001). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods, 6,* 352-370.

Enders, C. K. (2010). *Applied missing data analysis.* New York: Guilford Press.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 8,* 430-457.

Enders, C. K., & Gottschall, A. C. (2011). The impact of missing data on the ethical quality of a research study. In A. T. Panter & S. K. Sterba (Eds.), *Handbook of ethics in quantitative methodology* (pp. 357-381). New York, NY: Taylor & Francis.

Finney, S. J., Pieper, S. L., & Barron, K. E. (2004). Examining the psychometric properties of the Achievement Goal Questionnaire in a general academic context. *Educational and Psychological Measurement, 64,* 365-382.

Finney, S. J., Sundre, D. L., Swain, M., & Williams, L. M. (2014). Are value-added estimates influenced by test consequences in large-scale, low-stakes testing contexts? Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.

Ford, B. M. (1983). An overview of hot-deck procedures. In W. Madow, I. Olkin, & D. Rubin (Eds.), *Incomplete data in sample surveys, vol. 2* (pp. 185-207). New York: Academic Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple

sequences. *Statistical Science, 7*, 457-472.

Glynn, R. J., Laird, N. M., & Rubin, D. B. (1993). Multiple imputation in mixture models for nonignorable nonresponse with follow-ups. *Journal of the American Statistical Association, 88,* 984-993.

Gold, M. S., & Bentler, P. M. (2000). Treatments of missing data: A monte carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization. *Structural Equation Modeling: A Multidisciplinary Journal, 7*, 319-355.

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 10,* 80-100.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60,* 549-576.

Graham, J. W., & Donaldson, S. I. (1993). Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of follow-up data. *Journal of Applied Psychology, 78,* 119-128.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8,* 206-213.

Hansen, M. H., & Hurwiz, W. N. (1946). The problem of nonresponse in sample surveys. *Journal of the American Statistical Association, 41,* 517-529.

Heckman, J. T. (1976). The common structure of statistical models of truncation, sample

selection, and limited dependent variables and a simple estimator for such models. *The Annals of Economic and Social Measurement, 5,* 475-492.

Heckman, J. T. (1979). Sample selection bias as a specification error. *Econometrica, 47,* 153-161.

Heitjan, D. F., & Basu, S. (1996). Distinguishing "missing at random" and "missing completely at random". *The American Statistician, 50,* 207-213.

Horst, S. J. (2010). *A mixture-modeling approach to exploring test-taking motivation in large-scale low-stakes contexts* (Doctoral dissertation, James Madison University).

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3,* 424-453.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6,* 1-55.

Jackman, S. (2000). Estimation and inference via Bayesian simulation: An introduction to Markov chain Monte Carlo. *American Journal of Political Science, 44,* 375-404.

John, O. P., & Srivastava, S. (1999). The big-five trait taxonomy: History, measurement, and theoretical perspectives. In L. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research, 2nd ed* (pp. 102-138). New York: Guilford.

Jöreskog, K. G., & Sörbom, D. (1993). *PRELIS 2 user's reference guide* [Computer software]. Chicago: Scientific Software.

Keel, P. K., Mitchell, J. E., Davis, T. L., & Crow, S. J. (2002). Long-term impact of

treatment in women diagnosed with bulimia nervosa. *International Journal of Eating Disorders, 31,* 151-158.

Lavori, P. W. (1992). Clinical trials in psychiatry: Should protocol deviation censor patient data? *Neuropsychopharmacology, 6,* 39-48.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association, 90,* 1112-1121.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.

MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding, and suppression effect. *Prevention Science, 1,* 173-181.

Marsh, H. W., Hau, K-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal, 11,* 320-341.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29,* 67-101.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction.* New York, NY: Guildford Press.

Muthén, L. K., & Muthén, B. O. (1998-2013). *Mplus User's Guide. Seventh Edition.* Los Angeles, CA: Muthén & Muthén.

Olinski, A., Chen, S., & Harlow, L. (2003). The comparative efficacy of imputation

methods for missing data in structural equation modeling. *European Journal of Operational Research, 151,* 53-79.

Payne, S. C., Youngcourt, S. S., & Beaubien, J. M. (2007). A meta-analytic examination of the goal orientation nomological net. *Journal of Applied Psychology, 92,* 128-150.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74,* 525-556.

Pieper, S. L. (2003). *Refining and extending the 2 x 2 achievement goal framework: Another look at work-avoidance* (Doctoral dissertation, James Madison University).

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82,* 33-40.

Raymond, M. R., Neustel, S., & Anderson, D. (2009). Same-form retest effects on credentialing examinations. *Educational Measurement: Issues and Practice, 28*(2), 19-27.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581-592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* Hoboken, NJ: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91,* 473-489.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-

added assessment in education. *Journal of Educational and Behavioral Statistics, 29,* 103-116.

Savalei, V., & Bentler, P. M. (2009). A two-stage approach to missing data: Theory and application to auxiliary variables. *Structural Equation Modeling: A Multidisciplinary Journal, 16,* 477-497.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* Boca Raton, FL: Chapman & Hall.

Schafer, J. L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica, 57,* 19-35.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147-177.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33,* 545-571.

Schraw, G., & Dennison, R. S. (1994) Assessing metacognitive awareness. *Contemporary Educational Psychology, 19,* 460-475.

Sundre, D. L. (2008). *The Natural World Test, Version 9 (NW-9): A measure of quantitative and scientific reasoning: Test manual.* Harrisonburg, VA: Center for Assessment and Research Studies, James Madison University.

Sundre, D. L. & Wise, S. L. (2003, April). *'Motivation filtering': An exploration of the impact of low examinee motivation on the psychometric quality of tests.* Paper presented to the National Council on Measurement in Education. Chicago, IL.

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2009). Skipping the test: Using

empirical evidence to inform policy related to students who avoid taking low-stakes assessments in college. *Journal of General Education, 58,* 167-195.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association, 82,* 528-540.

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: Using the Student Opinion Scale to make valid inferences about student performance. *Journal of General Education, 58,* 129-151.

Thoemmes, F., & Rose, N. (in press). A cautious note on auxiliary variables that can increase bias in missing data problems. *Multivariate Behavioral Research.*

U.S. Department of Education (2006). *A test of leadership: Charting the future of U.S. higher education*. Washington, DC. Retrieved from http://www2.ed.gov/about/bdscomm/list/hiedfuture/reports/pre-pub-report.pdf

Wilkinson, L., & Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594-604.

Wise, S. L. (2009). Strategies for managing the problem of unmotivated examinees in low-stakes testing programs. *Journal of General Education, 58,* 152-165.

Wise, V. L., Wise, S. L., & Bhola, D. S. (2006). The generalizability of motivation filtering in improving test score validity. *Educational Assessment, 11*, 65-83.

Zilberberg, A. (2013). *Students' attitudes toward institutional accountability testing in higher education: Implications for the validity of test scores* (Doctoral dissertation, James Madison University).

Table 1
Missing Data Mechanisms

UNKNOWN		Missingness (*R*) related to measured variables (*X*)? YES	Missingness (*R*) related to measured variables (*X*)? NO
Missingness (*R*) related to variable with missing values (*Y*), after controlling for measured variables (*X*) included in the analysis?	YES	MNAR	MNAR
	NO	MAR	MCAR

Note. MCAR = Missing completely at random; MAR = Missing at random; MNAR = Missing not at random. The missing data mechanism underlying the data depends on whether missingness (*R*) is related to the variable with missingness itself (*Y*), related to other measured variables (*X*), and related to the variable with missingness itself (*Y*) conditional on the measured variables (*X*) included in the analysis. Typically, missingness variable *R* is computed by assigning a value of 0 for all cases where *Y* is missing, and a value of 1 for all cases where *Y* is observed. The researcher can empirically determine whether *R* is related to any measured variable if the measured variables are not missing for all cases where *Y* is missing. If a significant relationship exists between any measured variable and *R*, data are either MNAR or MAR, depending on whether *R* is related to *Y* after controlling for the measured variables. If a significant relationship does *not* exist between any measured variable and *R*, data are either MNAR or MCAR, again depending on whether *R* is related to *Y* after controlling for the measured variables. Unfortunately, the values of *Y* are always missing for all cases where $R = 0$, so the relationship between *R* and *Y* cannot be empirically estimated. Thus, MNAR data cannot be empirically differentiated from MCAR or MAR data in most missing data situations.

Table 2

Hypothesized Effects of Including Auxiliary Variables with Different Relationships with Missingness and Posttest Scores

	μ_y		σ^2_y		$cov_{x,y}$		μ_{y-x}	
Auxiliary variable relationships	Est.	SE	Est.	SE	Est.	SE	Est.	SE
Unrelated to posttest scores (*Y*)	No Change	No Change	No Change	No Change	No Change	No Change	No Change	No Change
Related to posttest scores (*Y*), unrelated to missingness (*R*)	No Change	Reduced	No Change	Reduced	No Change	Reduced	No Change	Reduced
Related to posttest scores (*Y*), linearly related to missingness (*R*)	Less Bias	Reduced	Less Bias	Reduced	Less Bias	Reduced	Less Bias	Reduced
Related to posttest scores (*Y*), nonlinearly related to missingness (*R*)	No change	Reduced	Less Bias	Reduced	Less Bias	Reduced	No Change	Reduced

Note. Affected parameters are highlighted in grey, based on research conducted by Collins and colleagues (2001). Including auxiliary variables unrelated to posttest scores will not result in improvement of parameter bias or standard errors. Including auxiliary variables unrelated to missingness but related to posttest scores will result in reduced standard errors, but no reduction in parameter bias. Including auxiliary variables *linearly* related to missingness and related to posttest scores will result in less parameter bias and reduced standard errors. Including auxiliary variables *nonlinearly* related to missingness and related to posttest scores will result in reduced standard errors, reduced *variance* and *covariance* parameter bias, and no change in *mean* and *mean difference* parameter bias.

Table 3

Methods for Dealing with Missingness

Missing Data Method	Description	Appropriate for Which Missing Data Mechanisms	Recommend?	Comments
Deletion Methods				
Listwise Deletion (LD)	Cases with missing data on any variables are deleted.	MCAR	Under MCAR conditions	LD will result in reduced power under MCAR conditions, but parameter estimates will be accurate.
Pairwise Deletion (PD)	If a case has missing data for a variable involved in a given parameter estimate, that case is excluded from estimating that parameter.	MCAR	No	PD can lead to significant model estimation problems due to nonpositive definite matrices.
Single Imputation Methods				
Mean Imputation	The mean for a variable is used to substitute for any missing values for that variable.	None	No	Mean imputation will always introduce bias and should never be used.

Table 3 (continued)

Methods for Dealing with Missingness

Missing Data Method	Description	Appropriate for Which Missing Data Mechanisms	Recommend?	Comments
Regression Imputation	Predicted values using a regression equation involving other dataset variables are used to substitute for missing values.	MCAR	No	Will only produce unbiased variance and covariance estimates under MCAR conditions when corrective adjustments are applied. Also, standard errors will be biased downward, and better techniques (MI, FIML) are now available.
Stochastic Regression Imputation	Similar to regression imputation, but a random error term is added when imputing missing values.	MCAR MAR	No	Standard errors will be biased downward, and better techniques (MI, FIML) are now available.
Modern Methods				
Multiple Imputation (MI)	Stochastic regression imputation is used to impute *multiple* datasets, and the variability in parameter estimates across those datasets is used in the calculation of standard errors for parameter estimates.	MCAR MAR	Under all conditions	20 imputations and a large number of iterations are generally recommended. Utilizing auxiliary variables can increase accuracy.

Table 3 (continued)

Methods for Dealing with Missingness

Missing Data Method	Description	Appropriate for Which Missing Data Mechanisms	Recommend?	Comments
Full information Maximum Likelihood (ML)	Available data used to estimate population parameter values that are most likely to have produced sample data (Baraldi & Enders, 2012).	MCAR MAR	Under all conditions	Utilizing auxiliary variables can increase accuracy.
MNAR-based methods (assorted)	Generally, the model of interest (e.g., growth model) is supplemented with an additional model of the probability of missingness.	MNAR	Under specific MNAR scenarios	Methods require strict *a priori* assumptions, and significant bias is introduced when these assumptions are not met. Thus, these methods are only recommended in very specific MNAR scenarios, where a strong theory of missingness is specified.

Table 4
Examined Auxiliary Variables

	Hypothesized to be Predictive of:			% Missingness MAP Sample		% Missingness NW-9 Sample	
Auxiliary Variable	Missingness	MAP Score	NW-9 Score	ADay	Makeup	ADay	Makeup
U. Database Variables							
Gender	X	X		0.8%	7.5%	0.7%	0.0%
Posttest Age	X			0.8%	0.0%	0.7%	0.0%
SAT Math		X	X	2.1%	3.7%	2.3%	2.2%
SAT Verbal		X	X	2.1%	3.7%	2.3%	2.2%
Posttest GPA	X			0.8%	0.0%	0.8%	0.0%
Posttest earned total credits	X			0.8%	0.0%	0.8%	0.0%
Posttest earned science credits			X	-	-	0.8%	0.0%
Pretest Variables							
MAP[a]	X	X		0.0%	0.0%	5.7%	4.3%
PAP	X	X		0.0%	0.0%	4.4%	4.3%
MAV		X		0.3%	1.5%	-	-
PAV		X		0.4%	0.0%	-	-
WAV	X	X		0.3%	0.7%	4.7%	4.3%
Openness		X		27.3%	23.9%	66.2%	63.0%
Conscientiousness	X	X		27.2%	23.1%	66.4%	60.9%
Extraversion		X		27.3%	23.1%	66.2%	60.9%
Agreeableness		X		27.4%	23.1%	66.3%	60.9%
Neuroticism		X		27.4%	24.6%	66.7%	63.0%
MAI-R[b]		X	X	1.5%	5.2%	4.4%	4.3%
Effort	X	X	X	23.5%	23.1%	25.0%	30.4%
Importance	X	X	X	11.7%	15.7%	15.8%	21.7%

Table 4 (continued)
Examined Auxiliary Variables

	Hypothesized to be Predictive of:			% Missingness MAP Sample		% Missingness NW-9 Sample	
Auxiliary Variable	Missingness	MAP Score	NW-9 Score	ADay	Makeup	ADay	Makeup
Posttest Variables							
MAP	X	N/A		0.0%	0.0%	0.1%	0.0%
PAP	X	X		0.1%	0.0%	0.3%	0.0%
MAV		X		0.6%	1.5%	0.9%	4.3%
PAV		X		0.2%	0.0%	0.4%	0.0%
WAV	X	X		0.3%	0.0%	0.8%	0.0%
Openness		X		53.7%	3.0%	49.2%	6.5%
Conscientiousness	X	X		53.4%	0.7%	49.3%	0.0%
Extraversion		X		53.3%	0.7%	49.0%	0.0%
Agreeableness		X		53.5%	2.2%	49.2%	0.0%
Neuroticism		X		53.3%	1.5%	49.2%	0.0%
Effort	X	X	X	1.2%	3.7%	1.1%	2.2%
Importance	X	X	X	0.6%	2.2%	0.8%	2.2%

Note. Due to students being randomly assigned to different testing configurations, missingness percentages vary across auxiliary variables. U. Database Variables = Variables obtained from the university student database; Pretest Variables = Variables measured at pretest for entering freshmen students; Posttest Variables = Variables measured at posttest after three semesters of university attendance; MAP = Mastery Approach Orientation; NW-9 = Natural World Version 9; PAP = Performance Approach Orientation; MAV = Mastery Avoidance Orientation; PAV = Performance Avoidance Orientation; WAV = Work Avoidance; MAI-R = Metacognitive Regulation; Effort = Test-taking Effort; Importance = Test-taking Importance.

[a] Although Pretest MAP is listed as an auxiliary variable and is hypothesized to be related to posttest MAP scores, pretest MAP was not considered a strictly auxiliary variable in the MAP analyses. That is, the Pretest MAP score was included as part of the MAP analysis model in computing difference scores for MAP growth estimates.

[b] Unlike the other auxiliary variables, Metacognitive Regulation was only measured at pretest.

Table 5

Descriptive Statistics and Model Parameters (Standard Errors) Regressing Posttest MAP Scores on Pretest MAP Scores by Posttest Attendance

Attendance	Pretest Mean	Pretest Variance	Posttest Mean	Posttest Variance
Assessment Day at Posttest	17.892	6.944	16.932	9.135
Makeup at Posttest	17.343	7.972	15.119	14.687

Attendance	Posttest Intercept	Pretest-Posttest Slope	Posttest Residual Variance
Assessment Day at Posttest	8.957 (0.415)	0.446 (0.023)	7.756 (0.238)
Makeup at Posttest	8.487 (1.977)	0.382 (0.113)	13.521 (1.652)

Note. Intercepts, slopes, and residual variances were freely estimated across groups. If students attending makeup testing responded comparably to students attending Assessment Day at posttest, we would expect these parameters to be of similar value, within sampling error. The makeup sample was associated with a smaller posttest mean and pretest-posttest slope, and a larger posttest variance and residual variance as compared to the Assessment Day sample. If models constraining common intercept, slope, and/or residual variance parameters across samples were associated with significant model misfit (see Table 6), makeup students may not be providing valid responses at posttest.

Table 6
Multiple Group Analysis Comparing the Pretest-Posttest MAP Relationship Across Assessment Day and Makeup Samples

Model	χ^2	*df*	CFI	RMSEA
Model 1: Posttest Intercept Constraint	0.054	1	>.999	<.001
Model 2: Pretest-Posttest Slope Constraint	0.303	1	>.999	<.001
Model 3: Posttest Residual Variance Constraint	22.992*	1	.938	.140
Model 4: Intercept, Slope, and Residual Variance Constraint	61.432*	3	.836	.131

Note. CFI= Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation. Models were estimated predicting posttest MAP scores from pretest MAP scores. When estimating Model 1, the posttest intercept was constrained to be equal across Assessment Day and makeup samples, but the pretest-posttest slopes and posttest residual variances were freely estimated across samples. When estimating Model 2, the pretest-posttest slope was constrained to be equal across samples, but the posttest intercept and posttest residual variances were freely estimated. When estimating Model 3, posttest residual variances are constrained to be equal across samples, but the posttest intercepts and pretest-posttest slopes were freely estimated. When estimating Model 4, the posttest intercept, pretest-posttest slope, and posttest residual variance are all constrained to be equal across samples. Results indicate that Models 3 and 4 are associated with statistically and practically significant misfit. The normalized posttest score variance residual associated with Model 4 was 2.701 for the makeup sample and -1.775 for the Assessment Day sample, indicating that the posttest score variance was underestimated by the model for the makeup sample. Additionally, the normalized posttest score mean residual associated with Model 4 was -4.450 for the makeup sample, indicating that the posttest score mean was overestimated by the model for the makeup sample. These results indicate that the posttest residual variance is not common across samples, with the makeup sample having a larger residual variance than the Assessment Day sample. This increased posttest residual variance may be due to reduced motivation by the makeup sample, resulting in increased random responding at posttest.
* $p < .05$.

Table 7

*Descriptive Statistics for the Complete MAP Sample (*N *=2254)*

Measure	Mean	SD	Min	Max
1. Posttest Attendance (*R*)	.941[a]	0.237	0.00	1.00
2. Posttest MAP Score (*Y*)	16.824	3.107	3.00	21.00
3. Pretest MAP Score	17.859	2.650	6.00	21.00
U. Database Variables				
4. Gender	.358[b]	0.479	0.00	1.00
5. Age	19.918	0.376	18.58	23.68
6. SAT Math	581.596	65.039	320.00	800.00
7. SAT Verbal	571.923	69.998	280.00	800.00
8. GPA	3.152	0.411	1.73	4.00
9. Posttest Credit Hours	51.805	5.975	45.00	70.00
Pretest Auxiliary Variables				
10. Pretest PAP	16.056	3.784	3.00	21.00
11. Pretest MAV	12.772	3.655	3.00	21.00
12. Pretest PAV	14.204	3.971	3.00	21.00
13. Pretest WAV	10.530	4.556	4.00	28.00
14. Pretest Openness	35.489	6.362	17.00	55.00
15. Pretest Conscientiousness	32.440	5.102	13.00	47.00
16. Pretest Extraversion	28.108	6.235	9.00	42.00
17. Pretest Agreeableness	36.025	4.968	18.00	50.00
18. Pretest Neuroticism	21.887	5.842	8.00	40.00
19. Pretest MAI-R	125.827	15.826	70.00	184.00
20. Pretest Effort	18.943	3.606	5.00	25.00
21. Pretest Importance	15.307	3.984	5.00	25.00
Posttest Auxiliary Variables				
22. Posttest PAP	15.794	4.064	3.00	21.00
23. Posttest MAV	26.374	6.152	6.00	42.00
24. Posttest PAV	13.745	4.031	3.00	21.00
25. Posttest WAV	12.029	4.933	4.00	28.00
26. Posttest Openness	37.120	6.307	15.00	50.00
27. Posttest Conscientiousness	33.433	5.461	12.00	45.00
28. Posttest Extraversion	28.568	6.290	10.00	40.00
29. Posttest Agreeableness	35.449	5.531	13.00	45.00
30. Posttest Neuroticism	22.342	6.057	8.00	40.00
31. Posttest Effort	18.991	3.698	5.00	25.00
32. Posttest Importance	13.622	4.430	5.00	25.00

Note. U. Database Variables = Variables obtained from the university student database; Pretest Auxiliary Variables = Variables measured at pretest for entering freshmen students; Posttest Auxiliary Variables = Variables measured at posttest after three semesters of university attendance; MAP = Mastery Approach Orientation; PAP = Performance Approach Orientation; MAV = Mastery Avoidance Orientation; PAV = Performance Avoidance Orientation; WAV = Work Avoidance; MAI-R = Metacognitive Regulation; Effort = Test-taking Effort; Importance = Test-taking Importance.

[a] This value (.941) represents the proportion of students attending their originally assigned Assessment Day testing session at posttest
[b] This value (.358) represents the proportion of males in the sample.

Table 8

Bivariate Relationships between Posttest Attendance (R), Posttest MAP Score (Y), and Potential Auxiliary Variables

Measure	1	2	3	4	5
1. Posttest Attendance (*R*)	---				
2. Posttest MAP Score (*Y*)	.138*	---			
3. Pretest MAP Score	.049*	.382*	---		
U. Database Variables					
4. Gender	-.083*	-.126*	-.155*	---	
5. Age	-.032	-.007	-.004	.157*	---
6. SAT Math	-.019	-.095*	-.172*	.307*	-.015
7. SAT Verbal	-.058*	-.092*	-.140*	.115*	-.044*
8. GPA	.068*	.086*	-.009	-.066*	-.011
9. Posttest Credit Hours	.005	.018	-.002	.075*	.000
Pretest Auxiliary Variables					
10. Pretest PAP	.016	.127*	.289*	.054*	.027
11. Pretest MAV	.049*	.073*	.217*	-.187*	-.039
12. Pretest PAV	.047*	.053*	.160*	-.146*	-.028
13. Pretest WAV	-.014	-.277*	-.460*	.164*	.006
14. Pretest Openness	-.056*	.113*	.175*	.070*	.044
15. Pretest Conscientiousness	.076*	.228*	.303*	-.186*	.013
16. Pretest Extraversion	-.022	.080*	.069*	-.115*	-.004
17. Pretest Agreeableness	.069*	.177*	.205*	-.182*	-.063*
18. Pretest Neuroticism	.004	.002	.026	-.242*	-.033
19. Pretest MAI-R	.026	.311*	.448*	-.127*	.021
20. Pretest Effort	.039	.149*	.167*	-.067*	-.040
21. Pretest Importance	.028	.162*	.164*	-.060*	.010
Posttest Auxiliary Variables					
22. Posttest PAP	.043*	.321*	.109*	.003	.006
23. Posttest MAV	.040	.480*	.242*	-.162*	-.044*
24. Posttest PAV	.058*	.134*	.089*	-.147*	-.034
25. Posttest WAV	-.073*	-.500*	-.258*	.162*	-.024
26. Posttest Openness	-.010	.208*	.151*	.032	.030
27. Posttest Conscientiousness	.099*	.321*	.228*	-.246*	-.042
28. Posttest Extraversion	-.014	.110*	.062*	-.135*	.014
29. Posttest Agreeableness	.110*	.260*	.176*	-.269*	-.107*
30. Posttest Neuroticism	.002	-.016	.047	-.246*	-.068*
31. Posttest Effort	.018	.236*	.099*	-.091*	-.002
32. Posttest Importance	.036	.216*	.067*	-.039	.005

Table 8 (continued)

Bivariate Relationships between Posttest Attendance (R), Posttest MAP Score (Y), and Potential Auxiliary Variables

Measure	6	7	8	9	10
U. Database Variables					
6. SAT Math	---				
7. SAT Verbal	.380*	---			
8. GPA	.204*	.260*	---		
9. Posttest Credit Hours	.172*	.257*	.185*	---	
Pretest Auxiliary Variables					
10. Pretest PAP	.020	-.039	.024	.033	---
11. Pretest MAV	-.130*	-.156*	-.026	-.089*	.125*
12. Pretest PAV	-.173*	-.216*	-.102*	-.092*	.391*
13. Pretest WAV	.148*	.129*	-.047*	.016	-.135*
14. Pretest Openness	.008	.258*	.043	.166*	.058*
15. Pretest Conscientiousness	-.136*	-.132*	.179*	.008	.185*
16. Pretest Extraversion	-.091*	-.093*	-.047	-.059*	.060*
17. Pretest Agreeableness	-.117*	-.178*	.004	-.054*	-.017
18. Pretest Neuroticism	-.154*	-.087*	.073*	-.039	-.018
19. Pretest MAI-R	-.141*	-.055*	.068*	.007	.252*
20. Pretest Effort	.050*	.096*	.121*	.090*	.084*
21. Pretest Importance	-.095*	-.080*	.024	-.012	.213*
Posttest Auxiliary Variables					
22. Posttest PAP	.028	-.035	.148*	.003	.473*
23. Posttest MAV	-.113*	-.080*	-.073*	-.019	.097*
24. Posttest PAV	-.160*	-.198*	-.152*	-.074*	.204*
25. Posttest WAV	.135*	.141*	-.030	.027	-.051*
26. Posttest Openness	.026*	.216*	.034	.127*	.054*
27. Posttest Conscientiousness	-.101*	-.109*	.192*	.008	.140*
28. Posttest Extraversion	-.119*	-.094*	-.083*	-.067*	.056*
29. Posttest Agreeableness	-.148*	-.170*	.023	-.066*	-.042
30. Posttest Neuroticism	-.113*	-.078*	.083*	-.033	.004
31. Posttest Effort	.068*	.063*	.097*	.041	.009
32. Posttest Importance	-.057*	-.062*	.013	-.030	.085*

Table 8 (continued)

Bivariate Relationships between Posttest Attendance (R), Posttest MAP Score (Y), and Potential Auxiliary Variables

Measure	11	12	13	14	15
Pretest Auxiliary Variables					
11. Pretest MAV	---				
12. Pretest PAV	.301*	---			
13. Pretest WAV	-.005	.013	---		
14. Pretest Openness	-.056*	-.039	-.108*	---	
15. Pretest Conscientiousness	-.008	.035	-.359*	.077*	---
16. Pretest Extraversion	-.016	.057*	-.088*	.202*	.116*
17. Pretest Agreeableness	.019	.065*	-.234*	.065*	.344*
18. Pretest Neuroticism	.215*	.109*	-.009	-.093*	-.103*
19. Pretest MAI-R	.088*	.128*	-.390*	.312*	.419*
20. Pretest Effort	-.044	-.002	-.144*	.115*	.230*
21. Pretest Importance	.044	.068*	-.169*	.068*	.173*
Posttest Auxiliary Variables					
22. Posttest PAP	.056*	.213*	-.072*	-.022	.149*
23. Posttest MAV	.267*	.195*	-.111*	.049*	.045
24. Posttest PAV	.183*	.437*	.005	-.044*	.008
25. Posttest WAV	.000	.005	.477*	-.053*	-.280*
26. Posttest Openness	-.057*	-.024	-.094*	.706*	.038
27. Posttest Conscientiousness	-.018	.028	-.266*	.004	.668*
28. Posttest Extraversion	-.024	.055*	-.087*	.165*	.119*
29. Posttest Agreeableness	.013	.048*	-.193*	.031	.268*
30. Posttest Neuroticism	.166*	.125*	-.057*	-.099*	-.042
31. Posttest Effort	-.027	-.046*	-.101*	.088*	.136*
32. Posttest Importance	.020	.031	-.093*	.026	.097*

Table 8 (continued)
Bivariate Relationships between Posttest Attendance (R), Posttest MAP Score (Y), and Potential Auxiliary Variables

Measure	16	17	18	19	20
Pretest Auxiliary Variables					
16. Pretest Extraversion	---				
17. Pretest Agreeableness	.180*	---			
18. Pretest Neuroticism	-.273*	-.260*	---		
19. Pretest MAI-R	.157*	.232*	-.054*	---	
20. Pretest Effort	.027	.142*	-.040	.202*	---
21. Pretest Importance	.021	.062*	.067*	.259*	.328*
Posttest Auxiliary Variables					
22. Posttest PAP	.040	.020	-.032	.164*	.072*
23. Posttest MAV	.044	.052*	.116*	.194*	.044
24. Posttest PAV	.004	.056*	.070*	.087*	.008
25. Posttest WAV	-.092*	-.193*	-.030	-.246*	-.116*
26. Posttest Openness	.181*	.068*	-.088*	.222*	.055
27. Posttest Conscientiousness	.117*	.251*	-.077*	.297*	.182*
28. Posttest Extraversion	.770*	.137*	-.193*	.134*	.024
29. Posttest Agreeableness	.109*	.675*	-.133*	.191*	.104*
30. Posttest Neuroticism	-.128*	-.137*	.660*	-.007	-.018
31. Posttest Effort	.035	.136*	-.015	.090*	.347*
32. Posttest Importance	.019	.066*	.029	.126*	.099*

Table 8 (continued)
Bivariate Relationships between Posttest Attendance (R), Posttest MAP Score (Y), and Potential Auxiliary Variables

Measure	21	22	23	24	25
Pretest Auxiliary Variables					
21. Pretest Importance	---				
Posttest Auxiliary Variables					
22. Posttest PAP	.153*	---			
23. Posttest MAV	.102*	.215*	---		
24. Posttest PAV	.078*	.459*	.410*	---	
25. Posttest WAV	-.105*	-.129*	-.182*	.009	---
26. Posttest Openness	.035	.044	.114*	-.031	-.116*
27. Posttest Conscientiousness	.095*	.233*	.075*	.026	-.377*
28. Posttest Extraversion	.021	.060*	.054	.019	-.113*
29. Posttest Agreeableness	.091*	.016	.124*	.086*	-.249*
30. Posttest Neuroticism	.031	.002	.100*	.095*	-.019
31. Posttest Effort	.116*	.113*	.090*	.022	-.208*
32. Posttest Importance	.372*	.142*	.131*	.049*	-.208*

Table 8 (continued)
Bivariate Relationships between Posttest Attendance (R), Posttest MAP Score (Y), and Potential Auxiliary Variables

Measure	26	27	28	29	30
Posttest Auxiliary Variables					
26. Posttest Openness	---				
27. Posttest Conscientiousness	.084*	---			
28. Posttest Extraversion	.240*	.178*	---		
29. Posttest Agreeableness	.140*	.379*	.169*	---	
30. Posttest Neuroticism	-.135*	-.091*	-.208*	-.199*	---
31. Posttest Effort	.181*	.205*	.067*	.218*	-.060*
32. Posttest Importance	.085*	.089*	.016	.106*	.008

Table 8 (continued)
Bivariate Relationships between Posttest Attendance (R), Posttest MAP Score (Y), and Potential Auxiliary Variables

Measure	31	32
Posttest Auxiliary Variables		
31. Posttest Effort	---	
32. Posttest Importance	.286*	---

Note. U. Database Variables = Variables obtained from the university student database; Pretest Auxiliary Variables = Variables measured at pretest for entering freshmen students; Posttest Auxiliary Variables = Variables measured at posttest after three semesters of university attendance; MAP = Mastery Approach Orientation; PAP = Performance Approach Orientation; MAV = Mastery Avoidance Orientation; PAV = Performance Avoidance Orientation; WAV = Work Avoidance; MAI-R = Metacognitive Regulation; Effort = Test-taking Effort; Importance = Test-taking Importance. Gender was coded 0 for female and 1 for male, and posttest attendance (*R*) was coded 0 for makeup and 1 for Assessment Day. Posttest attendance (*R*) was found to be significantly bivariately related posttest MAP scores (*Y*) as well as a number of other dataset variables (see column 1). Thus, the MCAR assumption was found to be violated.
* Sig. at $p < .05$.

Table 9
Partial Correlations between Posttest Attendance (R) and Posttest MAP Scores (Y) after Controlling for Individual Auxiliary Variables

Measure	X-R Cor.	X-Y Cor.	R-Y Partial	Partial - Bivariate
1. Pretest MAP Score	.049*	.382*	.129*	-.009
U. Database Variables				
2. Gender	-.083*	-.126*	.129*	-.009
3. Age	-.032	-.007	.138*	.000
4. SAT Math	-.019	-.095*	.137*	-.001
5. SAT Verbal	-.058*	-.092*	.133*	-.005
6. GPA	.068*	.086*	.133*	-.005
7. Posttest Credit Hours	.005	.018	.138*	.000
Pretest Auxiliary Variables				
8. Pretest PAP	.016	.127*	.137*	-.001
9. Pretest MAV	.049*	.073*	.135*	-.003
10. Pretest PAV	.047*	.053*	.136*	-.002
11. Pretest WAV	-.014	-.277*	.140*	.002
12. Pretest Openness	-.056*	.113*	.145*	.007
13. Pretest Conscientiousness	.076*	.228*	.124*	-.014
14. Pretest Extraversion	-.022	.080*	.140*	.002
15. Pretest Agreeableness	.069*	.177*	.128*	-.010
16. Pretest Neuroticism	.004	.002	.138*	.000
17. Pretest MAI-R	.026	.311*	.137*	-.001
18. Pretest Effort	.039	.149*	.134*	-.004
19. Pretest Importance	.028	.162*	.135*	-.003
Posttest Auxiliary Variables				
20. Posttest PAP	.043*	.321*	.131*	-.007
21. Posttest MAV	.040	.480*	.136*	-.002
22. Posttest PAV	.058*	.134*	.132*	-.006
23. Posttest WAV	-.073*	-.500*	.118*	-.020
24. Posttest Openness	-.010	.208*	.143*	.005
25. Posttest Conscientiousness	.099*	.321*	.113*	-.025
26. Posttest Extraversion	-.014	.110*	.140*	.002
27. Posttest Agreeableness	.110*	.260*	.114*	-.024
28. Posttest Neuroticism	.002	-.016	.138*	.000
29. Posttest Effort	.018	.236*	.138*	.000
30. Posttest Importance	.036	.216*	.133*	-.005

Note. The table above presents the bivariate correlation between each auxiliary variable and posttest attendance (X-R Cor.), the bivariate correlation between each auxiliary variable and posttest MAP score (X-Y Cor.), the partial correlation between posttest attendance and posttest MAP score after controlling for the given auxiliary variable (R-Y Partial), and the difference between the R-Y partial correlation and the R-Y bivariate correlation (Partial – Bivariate). Recall the bivariate relationship between R and Y equaled .138. Negative "Partial – Bivariate" values indicate that the given auxiliary variable (*X*) independently moderates the relationship between posttest attendance (*R*)

and posttest MAP scores (*Y*), and thus are important to include as auxiliary variables to reduce bias. The largest negative "Partial – Bivariate" values were associated with posttest auxiliary variables (WAV, Conscientiousness, and Agreeableness), suggesting that posttest auxiliary variables are important to include in order to minimize parameter bias.

* Sig. at $p < .05$.

Table 10
Model Comparison Predicting Posttest MAP Scores (Y) from Auxiliary Variables

Predictors Added to Model	R^2	R^2 Ch.	R-Y Partial
Model 1: + Pretest MAP Score	.146*	---	.129*
Model 2: + U. Database Variables	.162*	.016*	.116*
Model 3: + Pretest Aux. Variables	.198*	.036*	.119*
Model 4: + Posttest Aux. Variables	.526*	.328*	.108*

Note. R-Y Partial = Partial correlation between posttest attendance (*R*) and posttest MAP scores after controlling for variables included in the model. Recall the bivariate relationship between R and Y equaled .138. Each model includes all the predictors of the previous models, with additional predictors added. For example, Model 2 includes pretest MAP score and all university database variables as predictors of posttest MAP scores. R^2 and R^2 change significance was evaluated using Wald tests. The models indicate the variables that are significantly independently related to posttest MAP scores (*Y*), and were thus important to include as auxiliary variables to reduce standard errors. For example, the R^2 change associated with posttest auxiliary variables (Model 4) was .328 and statistically significant, indicating that the additional measured posttest variables were important to include as auxiliary variables to decrease standard errors. If the partial correlation was nonsignificant for a given model, the relationship between posttest attendance (*R*) and posttest MAP scores (*Y*) was completely moderated by the predictors in the model, indicating that the MAR assumption was met if these predictors were included as auxiliary variables. However, across Models 1 - 4, the partial correlation was significant, indicating a MNAR mechanism as missingness predicted a significant amount of variance in posttest scores after controlling for auxiliary variables.
* Sig. at $p < .05$.

Table 11

Regression Coefficients Predicting Posttest MAP Scores (Y) from Pretest MAP Scores, University Database Auxiliary Variables, and Pretest Auxiliary Variables

Predictor Variable	*B*	β	sr^2
1. Pretest MAP Score	0.303*	.259	.042
U. Database Variables			
2. Gender	-0.214	-.033	.001
3. Age	-0.012	-.001	<.001
4. SAT Math	0.000	-.003	<.001
5. SAT Verbal	-0.003*	-.057	.002
6. GPA	0.582*	.077	.005
7. Posttest Credit Hours	0.010	.019	<.001
Pretest Auxiliary Variables			
8. Pretest PAP	0.002	.002	<.001
9. Pretest MAV	0.000	-.001	<.001
10. Pretest PAV	-0.015	-.019	<.001
11. Pretest WAV	-0.042*	-.062	.003
12. Pretest Openness	0.010	.021	<.001
13. Pretest Conscientiousness	0.009	.015	<.001
14. Pretest Extraversion	0.010	.020	<.001
15. Pretest Agreeableness	0.032*	.052	.002
16. Pretest Neuroticism	0.003	.005	<.001
17. Pretest MAI-R	0.022*	.111	.008
18. Pretest Effort	0.030	.035	.001
19. Pretest Importance	0.042*	.054	.002

Note. Gender was coded 0 for female and 1 for male. b = unstandardized slope; β = standardized slope; sr^2 = squared semipartial correlation. Model R^2 = .198. Posttest auxiliary variables were excluded in this model, as they would not be available to assessment practitioners choosing to forgo makeup testing. Results including posttest auxiliary variables are included in Table 12. Results indicate that pretest MAP scores, some university database variables, and some pretest auxiliary variables were significant predictors of posttest MAP scores. Thus, these predictors were important to include as auxiliary variables in MI and FIML analyses to reduce standard errors.

* Sig. at $p < .05$

Table 12

Regression Coefficients Predicting Posttest MAP Scores (Y) from Pretest MAP Scores, University Database Auxiliary Variables, Pretest Auxiliary Variables, and Posttest Auxiliary Variables

Predictor Variable	*B*	β	sr^2
1. Pretest MAP Score	0.242*	.207	.026
U. Database Variables			
2. Gender	0.135	.021	<.001
3. Age	-0.014	-.002	<.001
4. SAT Math	-0.001	-.011	<.001
5. SAT Verbal	-0.002*	-.043	.001
6. GPA	0.380*	.050	.002
7. Posttest Credit Hours	0.009	.018	<.001
Pretest Auxiliary Variables			
8. Pretest PAP	-0.065*	-.079	.004
9. Pretest MAV	-0.038*	-.044	.002
10. Pretest PAV	-0.009	-.011	<.001
11. Pretest WAV	0.048*	.071	.003
12. Pretest Openness	-0.008	-.017	<.001
13. Pretest Conscientiousness	-0.024	-.040	.001
14. Pretest Extraversion	-0.009	-.018	<.001
15. Pretest Agreeableness	-0.007	-.011	<.001
16. Pretest Neuroticism	-0.007	-.013	<.001
17. Pretest MAI-R	0.010*	.052	.002
18. Pretest Effort	0.012	.014	<.001
19. Pretest Importance	0.015	.019	<.001
Posttest Auxiliary Variables			
20. Posttest PAP	0.183*	.240	.031
21. Posttest MAV	0.186*	.368	.096
22. Posttest PAV	-0.095*	-.123	.008
23. Posttest WAV	-0.199*	-.317	.064
24. Posttest Openness	0.035*	.072	.002
25. Posttest Conscientiousness	0.032	.056	.001
26. Posttest Extraversion	0.005	.009	<.001
27. Posttest Agreeableness	0.041*	.072	.002
28. Posttest Neuroticism	-0.007	-.013	<.001
29. Posttest Effort	0.037*	.045	.002
30. Posttest Importance	0.022	.031	.001

Note. Gender was coded (0 = female) (1 = male). b = unstandardized slope; β = standardized slope; sr^2 = squared semipartial correlation. Model R^2 = .526. Interestingly, the significant pretest auxiliary predictors of posttest MAP scores change when posttest auxiliary variables are included in the model. Thus, posttest auxiliary variables *moderate* the relationship between some pretest auxiliary variables (agreeableness and test-taking importance) and posttest MAP scores, and act as *suppressor variables* for some other pretest auxiliary variables (PAP and MAV).

* Sig. at $p < .05$

Table 13

Comparison of MAP Results Across Different Missing Data Handling Techniques

	Complete		Listwise		MI (no aux)		MI (U. vars and pretest aux only)		MI (all aux)	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
μ_y	16.824	0.065	16.932	0.066	16.914	0.066	16.906	0.066	16.884	0.064
sDiscrepancy or RE[a]	---	---	*1.662*	*1.031*	*1.385*	*1.031*	*1.262*	*1.031*	*0.923*	*0.969*
σ^2_y	9.649	0.287	9.135	0.281	9.149	0.280	9.162	0.278	9.191	0.282
sDiscrepancy or RE[a]	---	---	*-1.791*	*0.959*	*-1.742*	*0.952*	*-1.697*	*0.938*	*-1.596*	*0.965*
$cov_{x,y}$	3.148	0.186	3.095	0.186	3.116	0.187	3.144	0.185	3.141	0.185
sDiscrepancy or RE[a]	---	---	*-0.285*	*1.000*	*-0.172*	*1.011*	*-0.022*	*0.989*	*-0.038*	*0.989*
$\mu_{y\text{-}x}$	-1.035	0.068	-0.960	0.068	-0.945	0.069	-0.953	0.068	-0.975	0.067
sDiscrepancy or RE[a]	---	---	*1.103*	*1.000*	*1.324*	*1.030*	*1.206*	*1.000*	*0.882*	*0.971*

	FIML (no aux)		FIML (U. vars and pretest aux only)		FIML (all aux)	
	Est.	SE	Est.	SE	Est.	SE
μ_y	16.917	0.065	16.910	0.065	16.884	0.065
sDiscrepancy or RE[a]	*1.431*	*1.000*	*1.323*	*1.000*	*0.923*	*1.000*
σ^2_y	9.151	0.281	9.161	0.282	9.183	0.281
sDiscrepancy or RE[a]	*-1.735*	*0.959*	*-1.700*	*0.965*	*-1.624*	*0.959*
$cov_{x,y}$	3.130	0.186	3.143	0.186	3.143	0.184
sDiscrepancy or RE[a]	*-0.097*	*1.000*	*-0.027*	*1.000*	*-0.027*	*0.979*
$\mu_{y\text{-}x}$	-0.942	0.068	-0.948	0.068	-0.975	0.067
sDiscrepancy or RE[a]	*1.368*	*1.000*	*1.279*	*1.000*	*0.882*	*0.971*

Note. μ_y = mean posttest MAP score; σ^2_y = posttest MAP score variance; $cov_{x,y}$ = covariance between pretest and posttest MAP scores; $\mu_{y\text{-}x}$ = mean pre-post MAP score growth

[a] Standardized discrepancy (sDiscrepancy) is reported for parameter estimates, and relative efficiency (RE) is reported for standard errors. Standardized discrepancy quantifies the standard error difference between the parameter estimate obtained utilizing the missing data handling method and the complete data parameter estimate. Thus, standardized discrepancy values can be interpreted similarly to

z-scores, with values greater than ~|2| considered large and highlighted. Standardized discrepancy was negligible for all parameters across all methods. Relative efficiency quantifies the ratio between the squared standard errors obtained utilizing the missing data handling method and the squared standard errors obtained utilizing the complete dataset. Relative efficiency values can also be interpreted as the factor the sample size should be increased for a given missing data handling method to achieve the same standard errors as the complete dataset. For instance, the RE value for the listwise μ_y is 1.031, indicating that the listwise sample size should be increased by 3.1% to achieve the same μ_y standard error that was obtained using the complete dataset. Relative efficiency values greater than 1.2 were considered large and highlighted. No relative efficiency estimates indicated substantial standard error inflation.

Table 14

Comparison of MAP Results Across Different Missing Data Handling Techniques (25% Missingness)

	Complete		Listwise		MI (no aux)		MI (U. vars and pretest aux only)		MI (all aux)	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
μ_y	16.534	0.145	17.005	0.152	16.890	0.143	16.927	0.145	16.784	0.143
sDiscrepancy or RE[a]	---	---	*3.248*	*1.099*	*2.455*	*0.973*	*2.710*	*1.000*	*1.724*	*0.973*
σ^2_y	11.342	0.693	9.338	0.659	9.609	0.660	9.482	0.685	9.472	0.653
sDiscrepancy or RE[a]	---	---	*-2.892*	*0.904*	*-2.501*	*0.907*	*-2.684*	*0.977*	*-2.698*	*0.888*
$cov_{x,y}$	3.675	0.412	3.597	0.424	3.893	0.435	3.776	0.439	3.742	0.410
sDiscrepancy or RE[a]	---	---	*-0.189*	*1.059*	*0.529*	*1.115*	*0.245*	*1.135*	*0.163*	*0.990*
μ_{y-x}	-1.265	0.142	-0.945	0.145	-0.908	0.137	-0.871	0.140	-1.015	0.139
sDiscrepancy or RE[a]	---	---	*2.254*	*1.043*	*2.514*	*0.931*	*2.775*	*0.972*	*1.761*	*0.958*

	FIML (no aux)		FIML (U. vars and pretest aux only)		FIML (all aux)	
	Est.	SE	Est.	SE	Est.	SE
μ_y	16.919	0.149	16.908	0.150	16.785	0.144
sDiscrepancy or RE[a]	*2.655*	*1.056*	*2.579*	*1.070*	*1.731*	*0.986*
σ^2_y	9.489	0.674	9.502	0.673	9.533	0.653
sDiscrepancy or RE[a]	*-2.674*	*0.946*	*-2.655*	*0.943*	*-2.610*	*0.888*
$cov_{x,y}$	3.864	0.435	3.858	0.432	3.799	0.414
sDiscrepancy or RE[a]	*0.459*	*1.115*	*0.444*	*1.099*	*0.301*	*1.010*
μ_{y-x}	-0.879	0.144	-0.891	0.144	-1.014	0.139
sDiscrepancy or RE[a]	*2.718*	*1.028*	*2.634*	*1.028*	*1.768*	*0.958*

Note. μ_y = mean posttest MAP score; σ^2_y = posttest MAP score variance; $cov_{x,y}$ = covariance between pretest and posttest MAP scores; μ_{y-x} = mean pre-post MAP score growth

[a] Standardized discrepancy (sDiscrepancy) is reported for parameter estimates, and relative efficiency (RE) is reported for standard errors. Standardized discrepancy quantifies the standard error difference between the parameter estimate obtained utilizing the missing data handling method and the complete data parameter estimate. Thus, standardized discrepancy values can be interpreted similarly to

z-scores, with values greater than ~|2| considered large and highlighted. Standardized discrepancy estimates indicated that posttest variance estimates for all missing data handling techniques were substantially lower than those obtained using the complete dataset. Both posttest mean and pre-post mean change estimates for all missing data handling techniques were substantially higher than those obtained using the complete dataset, with the exception of MI and FIML estimation utilizing all auxiliary variables. Relative efficiency quantifies the ratio between the squared standard errors obtained utilizing the missing data handling method and the squared standard errors obtained utilizing the complete dataset. Relative efficiency values can also be interpreted as the factor the sample size should be increased for a given missing data handling method to achieve the same standard errors as the complete dataset. For instance, the RE value for the listwise μ_y is 1.099, indicating that the listwise sample size should be increased by 9.9% to achieve the same μ_y standard error that was obtained using the complete dataset. Relative efficiency values greater than 1.2 were considered large and highlighted. No relative efficiency estimates indicated substantial standard error inflation.

Table 15

Comparison of MAP Results Across Different Missing Data Handling Techniques (50% Missingness)

	Complete		Listwise		MI (no aux)		MI (U. vars and pretest aux only)		MI (all aux)	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
μ_y	16.052	0.220	16.985	0.267	16.877	0.238	16.846	0.244	16.472	0.244
sDiscrepancy or RE[a]	---	---	*4.241*	*1.473*	*3.750*	*1.170*	*3.609*	*1.230*	*1.909*	*1.230*
σ^2_y	12.997	1.123	9.567	1.169	9.429	1.236	9.754	1.295	11.261	1.331
sDiscrepancy or RE[a]	---	---	*-3.054*	*1.084*	*-3.177*	*1.211*	*-2.888*	*1.330*	*-1.546*	*1.405*
$cov_{x,y}$	4.063	0.682	4.749	0.884	4.586	0.782	4.532	0.790	4.614	0.729
sDiscrepancy or RE[a]	---	---	*1.006*	*1.680*	*0.767*	*1.315*	*0.688*	*1.342*	*0.808*	*1.143*
$\mu_{y\text{-}x}$	-1.466	0.222	-0.709	0.254	-0.642	0.231	-0.673	0.238	-1.046	0.237
sDiscrepancy or RE[a]	---	---	*3.410*	*1.309*	*3.712*	*1.083*	*3.572*	*1.149*	*1.892*	*1.140*

	FIML (no aux)		FIML (U. vars and pretest aux only)		FIML (all aux)	
	Est.	SE	Est.	SE	Est.	SE
μ_y	16.888	0.248	16.863	0.250	16.522	0.244
sDiscrepancy or RE[a]	*3.800*	*1.271*	*3.686*	*1.291*	*2.136*	*1.230*
σ^2_y	9.483	1.130	9.646	1.147	10.469	1.272
sDiscrepancy or RE[a]	*-3.129*	*1.013*	*-2.984*	*1.043*	*-2.251*	*1.283*
$cov_{x,y}$	4.596	0.758	4.574	0.742	4.628	0.728
sDiscrepancy or RE[a]	*0.782*	*1.235*	*0.749*	*1.184*	*0.828*	*1.139*
$\mu_{y\text{-}x}$	-0.631	0.241	-0.655	0.244	-0.997	0.236
sDiscrepancy or RE[a]	*3.761*	*1.178*	*3.653*	*1.208*	*2.113*	*1.130*

Note. μ_y = mean posttest MAP score; σ^2_y = posttest MAP score variance; $cov_{x,y}$ = covariance between pretest and posttest MAP scores; $\mu_{y\text{-}x}$ = mean pre-post MAP score growth

[a] Standardized discrepancy (sDiscrepancy) is reported for parameter estimates, and relative efficiency (RE) is reported for standard errors. Standardized discrepancy quantifies the standard error difference between the parameter estimate obtained utilizing the missing data handling method and the complete data parameter estimate. Thus, standardized discrepancy values can be interpreted similarly to

z-scores, with values greater than ~|2| considered large and highlighted. With the exception of the parameter estimates obtained utilizing MI with all auxiliary variables, all missing data techniques resulted in posttest mean and pre-post mean change estimates that were substantially higher than those obtained utilizing the complete data, and posttest variance estimates that were substantially lower. Relative efficiency quantifies the ratio between the squared standard errors obtained utilizing the missing data handling method and the squared standard errors obtained utilizing the complete dataset. Relative efficiency values can also be interpreted as the factor the sample size should be increased for a given missing data handling method to achieve the same standard errors as the complete dataset. For instance, the RE value for the listwise μ_y is 1.473, indicating that the listwise sample size should be increased by 47.3% to achieve the same μ_y standard error that was obtained using the complete dataset. Relative efficiency values greater than 1.2 were considered large and highlighted. Generally, standard error inflation was problematic across techniques, although results were inconsistent.

Table 16

Descriptive Statistics and Model Parameters (Standard Errors) Regressing Posttest NW-9 Scores on Pretest NW-9 Scores by Posttest Attendance

Attendance	Pretest Mean	Pretest Variance	Posttest Mean	Posttest Variance
Assessment Day at Posttest	44.075	55.907	48.833	54.918
Makeup at Posttest	45.500	60.685	46.870	85.809

Attendance	Posttest Intercept	Pretest-Posttest Slope	Posttest Residual Variance
Assessment Day at Posttest	19.807 (1.179)	0.659 (0.026)	30.670 (1.544)
Makeup at Posttest	8.437 (5.696)	0.845 (0.123)	42.507 (8.863)

Note. Intercepts, slopes, and residual variances were freely estimated across groups. If students attending makeup testing responded comparably to students attending Assessment Day at posttest, we would expect these parameters to be of similar value, within sampling error. The makeup sample was associated with a smaller posttest intercept, a larger pretest-posttest slope, and a larger posttest residual variance as compared to the Assessment Day sample. If models constraining common intercept, slope, and/or residual variance parameters across samples were associated with significant model misfit (see Table 17), makeup students may not be providing valid responses at posttest.

Table 17
Multiple Group Analysis Comparing the Pretest-Posttest NW-9 Relationship across Assessment Day and Makeup Samples

Model	χ^2	*df*	CFI	RMSEA
Model 1: Posttest Intercept Constraint	3.682	1	.995	.080
Model 2: Pretest-Posttest Slope Constraint	2.130	1	.998	.052
Model 3: Posttest Residual Variance Constraint	2.554	1	.997	.061
Model 4: Intercept, Slope, and Residual Variance Constraint	17.121*	3	.971	.106

Note. CFI= Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation. Models were estimated predicting posttest NW-9 scores from pretest NW-9 scores. When estimating Model 1, the posttest intercept was constrained to be equal across Assessment Day and makeup samples, but the pretest-posttest slopes and posttest residual variances were freely estimated across samples. When estimating Model 2, the pretest-posttest slope was constrained to be equal across samples, but the posttest intercept and posttest residual variances were freely estimated. When estimating Model 3, posttest residual variances were constrained to be equal across samples, but the posttest intercepts and pretest-posttest slopes were freely estimated. When estimating Model 4, the posttest intercept, pretest-posttest slope, and posttest residual variance were all constrained to be equal across samples. These global fit indices indicated that Model 4 was associated with statistically and practically significant misfit. However, no normalized residual variances or covariances associated with Model 4 were greater than |2| for either sample. Yet, the normalized posttest mean residual was large for the makeup (-2.015) sample, indicating that posttest NW-9 scores were lower for the makeup sample than would be predicted given their pretest scores, manifesting in a lower intercept. This lower intercept may be due to reduced motivation by the makeup sample, resulting in increased random responding at posttest leading to lower posttest scores.
* $p < .05$.

Table 18

Descriptive Statistics for the Complete NW-9 Sample (N = 835)

Measure	Mean	SD	Min	Max
1. Posttest Attendance (*R*)	.945[a]	-	-	-
2. Posttest NW-9 Score (*Y*)	48.725	7.542	18.00	66.00
3. Pretest NW-9 Score	44.153	7.506	17.00	65.00
U. Database Variables				
4. Gender	.357[b]	-	-	-
5. Age	19.932	0.369	18.70	22.49
6. SAT Math	578.151	65.687	380.00	750.00
7. SAT Verbal	572.190	73.155	280.00	800.00
8. GPA	3.164	0.404	1.83	4.00
9. Posttest Credit Hours	52.313	6.133	45.00	70.00
10. Posttest Science Credit Hours	7.327	3.943	0.00	23.00
Pretest Auxiliary Variables				
11. Pretest MAP	17.788	2.783	6.00	21.00
12. Pretest PAP	15.952	3.829	3.00	21.00
13. Pretest WAV	10.478	4.530	4.00	26.00
14. Pretest Conscientiousness	32.052	5.103	18.00	44.00
15. Pretest MAI-R	126.170	15.710	78.00	174.00
16. Pretest Effort	18.719	3.565	5.00	25.00
17. Pretest Importance	15.302	3.987	5.00	25.00
Posttest Auxiliary Variables				
18. Posttest MAP	16.724	3.222	3.00	21.00
19. Posttest PAP	15.568	4.222	3.00	21.00
20. Posttest WAV	12.087	4.949	4.00	28.00
21. Posttest Conscientiousness	33.198	5.333	12.00	45.00
22. Posttest Effort	19.147	3.604	5.00	25.00
23. Posttest Importance	13.782	4.426	5.00	25.00

Note. U. Database Variables = Variables obtained from the university student database; Pretest Auxiliary Variables = Variables measured at pretest for entering freshmen students; Posttest Auxiliary Variables = Variables measured at posttest after three semesters of university attendance; NW-9 = Natural World Version 9; MAP = Mastery Approach Orientation; PAP = Performance Approach Orientation; WAV = Work Avoidance; MAI-R = Metacognitive Regulation; Effort = Test-taking Effort; Importance = Test-taking Importance.

[a] This value (.945) represents the proportion of students attending their originally assigned Assessment Day testing session at posttest

[b] This value (.357) represents the proportion of males in the sample.

Table 19

Bivariate Relationships between Posttest Attendance (R), Posttest NW-9 Score (Y), and Potential Auxiliary Variables

Measure	1	2	3	4	5	6	7	8
1. Posttest Attendance (*R*)	---							
2. Posttest NW-9 Score (*Y*)	.059	---						
3. Pretest NW-9 Score	-.043	.663*	---					
U. Database Variables								
4. Gender	-.105*	.169*	.178*	---				
5. Age	.007	-.016	-.014	.135*	---			
6. SAT Math	-.064	.428*	.409*	.291*	.016	---		
7. SAT Verbal	-.081*	.536*	.516*	.123*	-.003	.367*	---	
8. GPA	.025	.344*	.295*	-.044	.005	.260*	.335*	---
9. Posttest Credit Hours	-.007	.217*	.299*	.049	-.027	.162*	.256*	.203*
10. Posttest Science Credit Hours	.033	.138*	.099*	.015	-.007	.087*	-.028	-.037
Pretest Auxiliary Variables								
11. Pretest MAP	.079*	-.030	-.044	-.168*	-.016	-.161*	-.101*	.039
12. Pretest PAP	-.029	.020	.021	.050	.043	.056	-.003	.100*
13. Pretest WAV	-.029	.068	.028	.125*	.034	.118*	.101*	-.053
14. Pretest Conscientiousness	.091*	.007	.003	-.174*	.001	-.175*	-.133*	.240*
15. Pretest MAI-R	.075*	-.050	-.024	-.135*	.054	-.121*	-.085*	.082*
16. Pretest Effort	.005	.213*	.319*	-.130*	-.062	.066	.108*	.165*
17. Pretest Importance	.008	-.047	.030	-.093*	-.004	-.116*	-.068	.065
Posttest Auxiliary Variables								
18. Posttest MAP	.136*	.016	-.085*	-.107*	-.021	-.076*	-.085*	.096*
19. Posttest PAP	.005	-.004	-.007	.024	-.014	.030	-.027	.136*
20. Posttest WAV	-.129*	.047	.094*	.158*	-.007	.115*	.153*	-.004
21. Posttest Conscientiousness	.061	.086	.012	-.152*	.002	-.073	-.093*	.200*
22. Posttest Effort	.032	.187*	.112*	-.131*	.009	.034	.016	.083*
23. Posttest Importance	.037	.000	-.083*	.010	.036	-.035	-.069*	.030

Table 19 (continued)

Bivariate Relationships between Posttest Attendance (R), Posttest NW-9 Score (Y), and Potential Auxiliary Variables

Measure	9	10	11	12	13	14	15	16
U. Database Variables								
9. Posttest Credit Hours	---							
10. Posttest Science Credit Hours	.180*	---						
Pretest Auxiliary Variables								
11. Pretest MAP	.031	.040	---					
12. Pretest PAP	.044	.027	.297*	---				
13. Pretest WAV	-.050	-.029	-.500*	-.152*	---			
14. Pretest Conscientiousness	.056	.007	.354*	.258*	-.388*	---		
15. Pretest MAI-R	.009	.055	.470*	.269*	-.355*	.442*	---	
16. Pretest Effort	.154*	.090*	.235*	.133*	-.196*	.377*	.252*	---
17. Pretest Importance	.012	.058	.180*	.262*	-.178*	.272*	.271*	.288*
Posttest Auxiliary Variables								
18. Posttest MAP	.003	.077*	.356*	.121*	-.281*	.249*	.346*	.169*
19. Posttest PAP	-.007	.114*	.117*	.456*	-.085*	.174*	.179*	.095*
20. Posttest WAV	.030	-.055	-.283*	-.043	.472*	-.279*	-.240*	-.112*
21. Posttest Conscientiousness	.103*	.040	.182*	.150*	-.235*	.676*	.217*	.239*
22. Posttest Effort	.049	.025	.167*	.061	-.144*	.228*	.138*	.378*
23. Posttest Importance	-.046	.052	.139*	.120*	-.117*	.125*	.186*	.015

Table 19 (continued)

Bivariate Relationships between Posttest Attendance (R), Posttest NW-9 Score (Y), and Potential Auxiliary Variables

Measure	17	18	19	20	21	22
Pretest Auxiliary Variables						
17. Pretest Importance	---					
Posttest Auxiliary Variables						
18. Posttest MAP	.198*	---				
19. Posttest PAP	.209*	.348*	---			
20. Posttest WAV	-.163*	-.500*	-.124*	---		
21. Posttest Conscientiousness	.172*	.320*	.226*	-.369*	---	
22. Posttest Effort	.147*	.262*	.113*	-.218*	.262*	---
23. Posttest Importance	.295*	.259*	.137*	-.232*	.088*	.213*

Note. U. Database Variables = Variables obtained from the university student database; Pretest Auxiliary Variables = Variables measured at pretest for entering freshmen students; Posttest Auxiliary Variables = Variables measured at posttest after three semesters of university attendance; NW-9 = Natural World Version 9; MAP = Mastery Approach Orientation; PAP = Performance Approach Orientation; WAV = Work Avoidance; MAI-R = Metacognitive Regulation; Effort = Test-taking Effort; Importance = Test-taking Importance. Gender was coded 0 for female and 1 for male, and posttest attendance (*R*) was coded 0 for makeup and 1 for Assessment Day. Posttest attendance (*R*) was found to be significantly bivariately related to a number of dataset variables (see column 1). Thus, the MCAR assumption was found to be violated.

* Sig. at $p < .05$.

Table 20

Partial Correlations between Posttest Attendance (R) and Posttest NW-9 Scores (Y) after Controlling for Individual Auxiliary Variables

Measure	X-R Cor.	X-Y Cor.	R-Y Partial	Partial - Bivariate
1. Pretest NW-9 Score	-.043	.663*	.117*	.058
U. Database Variables				
2. Gender	-.105*	.169*	.078*	.019
3. Age	.007	-.016	.059	.000
4. SAT Math	-.064	.428*	.096*	.037
5. SAT Verbal	-.081*	.536*	.122*	.063
6. GPA	.025	.344*	.054	-.005
7. Posttest Credit Hours	-.007	.217*	.062	.003
8. Posttest Science Credit Hours	.033	.138*	.055	-.004
Pretest Auxiliary Variables				
9. Pretest MAP	.079*	-.030	.062	.003
10. Pretest PAP	-.029	.020	.060	.001
11. Pretest WAV	-.029	.068	.061	.002
12. Pretest Conscientiousness	.091*	.007	.059	.000
13. Pretest MAI-R	.075*	-.050	.063	.004
14. Pretest Effort	.005	.213*	.059	.000
15. Pretest Importance	.008	-.047	.059	.000
Posttest Auxiliary Variables				
16. Posttest MAP	.136*	.016	.057	-.002
17. Posttest PAP	.005	-.004	.059	.000
18. Posttest WAV	-.129*	.047	.066	.007
19. Posttest Conscientiousness	.061	.086	.054	-.005
20. Posttest Effort	.032	.187*	.054	-.005
21. Posttest Importance	.037	.000	.059	.000

Note. The table above presents the bivariate correlation between each auxiliary variable and posttest attendance (X-R Cor.), the bivariate correlation between each auxiliary variable and posttest NW-9 score (X-Y Cor.), the partial correlation between posttest attendance and posttest NW-9 score after controlling for the given auxiliary variable (R-Y Partial), and the difference between the R-Y partial correlation and the R-Y bivariate correlation (Partial – Bivariate). Recall the bivariate relationship between R and Y equaled .059. Negative "Partial – Bivariate" values indicate that the given auxiliary variable (*X*) independently moderates the relationship between posttest attendance (*R*) and posttest NW-9 scores (*Y*), and thus are important to include as auxiliary variables to reduce bias. In contrast to the MAP results, there are several auxiliary variables (*X*) with large *positive* "Partial – Bivariate" values, indicating that the partial correlation between posttest attendance (*R*) and posttest NW-9 scores (*Y*) *increases* when the auxiliary variable (*X*) is accounted for (pretest NW-9 score, Gender, SAT Math, SAT Verbal). These variables are examples of *suppressor variables*, as accounting for these variables increases the R-Y relationship.

* Sig. at $p < .05$

Table 21
Model Comparison Predicting Posttest NW-9 Scores (Y) from Auxiliary Variables

Predictors Added to Model	R^2	R^2 Ch.	R-Y Partial
Model 1: + Pretest NW-9 Score	.440*	---	.117*
Model 2: + U. Database Variables	.527*	.087*	.149*
Model 3: + Pretest Aux. Variables	.536*	.009	.143*
Model 4: + Posttest Aux. Variables	.562*	.026*	.139*

Note. R-Y Partial = Partial correlation between posttest attendance (*R*) and posttest NW-9 scores (*Y*) after controlling for variables included in the model. Recall the bivariate relationship between R and Y equaled .059. Each model included all the predictors of the previous models, with additional predictors added. For example, Model 2 included pretest NW-9 score and all university database variables as predictors of posttest NW-9 score. R^2 and R^2 change significance were evaluated using Wald tests. The results provide some indication of the *sets* variables that are significantly independent related to posttest NW-9 scores (*Y*), and were thus important to include as auxiliary variables to reduce standard errors. For example, the R^2 change associated with university database variables was .087 and statistically significant, indicating that university database variables were important to include as auxiliary variables to decrease standard errors. If the partial correlation was nonsignificant for a given model, the relationship between posttest attendance (*R*) and posttest NW-9 scores (*Y*) was completely moderated by the predictors in the model, indicating the MAR assumption was met if these predictors were included as auxiliary variables. Notice that for Models 1-4, the partial correlation was significant, indicating a MNAR mechanism as missingness predicted a significant amount of variance in posttest scores after controlling for auxiliary variables. Moreover, the partial correlation *increases* above the R-Y bivariate correlation (r = .059) when auxiliary variables are included due to *statistical suppression*.
* Sig. at $p < .05$.

Table 22

Regression Coefficients Predicting Posttest NW-9 Scores (Y) from Pretest NW-9 Scores, University Database Auxiliary Variables, and Pretest Auxiliary Variables

Predictor Variable	*b*	β	sr^2
1. Pretest NW-9 Score	0.457*	.455	.120
U. Database Variables			
2. Gender	0.633	.040	.001
3. Age	-0.292	-.014	<.001
4. SAT Math	0.014*	.122	.010
5. SAT Verbal	0.024*	.233	.034
6. GPA	1.963*	.105	.008
7. Posttest Credit Hours	-0.055	-.045	.002
8. Posttest Science Credit Hours	0.197*	.103	.010
Pretest Auxiliary Variables			
9. Pretest MAP	0.177*	.065	.003
10. Pretest PAP	-0.031	-.016	<.001
11. Pretest WAV	0.095	.057	.002
12. Pretest Conscientiousness	0.104	.071	.003
13. Pretest MAI-R	-0.019	-.040	.001
14. Pretest Effort	0.031	.015	<.001
15. Pretest Importance	-0.094	-.050	.002

Note. Gender was coded 0 for female and 1 for male. *b* = unstandardized slope; β = standardized slope; sr^2 = squared semipartial correlation. Model R^2 = .536. Posttest auxiliary variables were excluded in this model, as they would not be available to assessment practitioners choosing to forgo makeup testing. Results including posttest auxiliary variables are included in Table 23. Results indicate that pretest NW-9 score, some university database variables, and some pretest auxiliary variables were important predictors of posttest NW-9 scores. Thus, these predictors were important to include as auxiliary variables in MI and FIML analyses to reduce standard errors.

* $p < .05$

Table 23

Regression Coefficients Predicting Posttest NW-9 Scores (Y) from Pretest NW-9 Scores, University Database Auxiliary Variables, Pretest Auxiliary Variables, and Posttest Auxiliary Variables

Predictor Variable	*b*	β	sr^2
1. Pretest NW-9 Score	0.464*	.461	.121
U. Database Variables			
2. Gender	0.888*	.056	.003
3. Age	-0.473	-.023	.001
4. SAT Math	0.012*	.104	.007
5. SAT Verbal	0.024*	.237	.035
6. GPA	1.976*	.106	.008
7. Posttest Credit Hours	-0.064*	-.052	.002
8. Posttest Science Credit Hours	0.199*	.104	.010
Pretest Auxiliary Variables			
9. Pretest MAP	0.117	.043	.001
10. Pretest PAP	0.039	.020	<.001
11. Pretest WAV	0.105*	.063	.002
12. Pretest Conscientiousness	-0.021	-.014	<.001
13. Pretest MAI-R	-0.018	-.037	.001
14. Pretest Effort	-0.035	-.017	<.001
15. Pretest Importance	-0.122*	-.064	.003
Posttest Auxiliary Variables			
16. Posttest MAP	0.135	.058	.002
17. Posttest PAP	-0.127*	-.071	.003
18. Posttest WAV	0.004	.003	<.001
19. Posttest Conscientiousness	0.147	.103	.005
20. Posttest Effort	0.231*	.111	.009
21. Posttest Importance	0.062	.036	.001

Note. Gender was coded 0 for female and 1 for male. *b* = unstandardized slope; β = standardized slope; sr^2 = squared semipartial correlation. Model R^2 = .562. Note pretest importance was not a statistically significant predictor when posttest auxiliary variables were excluded from the model, but became significant after posttest auxiliary variables were included in the model. Thus, the posttest auxiliary variables acted as suppressor variables for pretest importance scores in the model. Results indicate that pretest NW-9 score, some university database variables, and some pretest and posttest auxiliary variables were important predictors of posttest NW-9 scores. Thus, these predictors were important to include as auxiliary variables in MI and FIML analyses to reduce standard errors.

* $p < .05$

Table 24

Comparison of NW-9 Results Across Different Missing Data Handling Techniques

	Complete		Listwise		MI (no aux)		MI (U. vars and pretest aux only)		MI (all aux)	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
μ_y	48.725	0.261	48.833	0.264	48.884	0.260	48.903	0.261	48.894	0.260
sDiscrepancy or RE[a]	---	---	*0.414*	*1.023*	*0.609*	*0.992*	*0.682*	*1.000*	*0.648*	*0.992*
σ^2_y	56.820	2.781	54.915	2.765	55.119	2.753	55.124	2.767	55.022	2.730
sDiscrepancy or RE[a]	---	---	*-0.685*	*0.989*	*-0.612*	*0.980*	*-0.610*	*0.990*	*-0.647*	*0.964*
$cov_{x,y}$	37.469	2.347	36.817	2.368	37.037	2.329	37.054	2.347	36.930	2.321
sDiscrepancy or RE[a]	---	---	*-0.278*	*1.018*	*-0.184*	*0.985*	*-0.177*	*1.000*	*-0.230*	*0.978*
μ_{y-x}	4.571	0.214	4.758	0.217	4.731	0.215	4.750	0.217	4.740	0.216
sDiscrepancy or RE[a]	---	---	*0.874*	*1.028*	*0.748*	*1.009*	*0.836*	*1.028*	*0.790*	*1.019*

	FIML (no aux)		FIML (U. vars and pretest aux only)		FIML (all aux)	
	Est.	SE	Est.	SE	Est.	SE
μ_y	48.884	0.261	48.914	0.260	48.904	0.260
sDiscrepancy or RE[a]	*0.609*	*1.000*	*0.724*	*0.992*	*0.686*	*0.992*
σ^2_y	55.078	2.763	54.984	2.748	54.924	2.742
sDiscrepancy or RE[a]	*-0.626*	*0.987*	*-0.660*	*0.976*	*-0.682*	*0.972*
$cov_{x,y}$	37.062	2.343	36.995	2.336	36.942	2.333
sDiscrepancy or RE[a]	*-0.173*	*0.997*	*-0.202*	*0.991*	*-0.225*	*0.988*
μ_{y-x}	4.731	0.216	4.761	0.216	4.751	0.215
sDiscrepancy or RE[a]	*0.748*	*1.019*	*0.888*	*1.019*	*0.841*	*1.009*

Note. μ_y = mean posttest NW-9 score; σ^2_y = posttest NW-9 score variance; $cov_{x,y}$ = covariance between pretest and posttest NW-9 scores; μ_{y-x} = mean pre-post NW-9 score growth

[a] Standardized discrepancy (sDiscrepancy) is reported for parameter estimates, and relative efficiency (RE) is reported for standard errors. Standardized discrepancy quantifies the standard error difference between the parameter estimate obtained utilizing the missing data handling method and the complete data parameter estimate. Thus, standardized discrepancy values can be interpreted similarly to

z-scores, with values greater than ~|2| considered large. Standardized discrepancy was negligible for all parameters across all methods. Relative efficiency quantifies the ratio between the squared standard errors obtained utilizing the missing data handling method and the squared standard errors obtained utilizing the complete dataset. Relative efficiency values can also be interpreted as the factor the sample size should be increased for a given missing data handling method to achieve the same standard errors as the complete dataset. For instance, the RE value for the listwise μ_y is 1.023, indicating that the listwise sample size should be increased by 2.3% to achieve the same μ_y standard error that was obtained using the complete dataset. Relative efficiency values greater than 1.2 were considered large. No relative efficiency estimates indicated substantial standard error inflation.

Table 25

Comparison of NW-9 Results Across Different Missing Data Handling Techniques (25% Missingness)

	Complete		Listwise		MI (no aux)		MI (U. vars and pretest aux only)		MI (all aux)	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
μ_y	48.228	0.596	48.681	0.647	48.856	0.623	49.002	0.584	49.003	0.603
sDiscrepancy or RE[a]	---	---	*0.760*	*1.178*	*1.054*	*1.093*	*1.299*	*0.960*	*1.300*	*1.024*
σ^2_y	65.328	6.811	57.681	6.944	59.567	7.278	58.214	7.032	58.249	6.505
sDiscrepancy or RE[a]	---	---	*-1.123*	*1.039*	*-0.846*	*1.142*	*-1.044*	*1.066*	*-1.039*	*0.912*
$cov_{x,y}$	38.168	5.224	34.234	5.509	36.270	5.419	34.429	5.139	34.252	5.206
sDiscrepancy or RE[a]	---	---	*-0.753*	*1.112*	*-0.363*	*1.076*	*-0.716*	*0.968*	*-0.750*	*0.993*
$\mu_{y\text{-}x}$	3.440	0.487	4.130	0.548	4.068	0.539	4.214	0.513	4.215	0.536
sDiscrepancy or RE[a]	---	---	*1.417*	*1.266*	*1.290*	*1.225*	*1.589*	*1.110*	*1.591*	*1.211*

	FIML (no aux)		FIML (U. vars and pretest aux only)		FIML (all aux)	
	Est.	SE	Est.	SE	Est.	SE
μ_y	48.837	0.619	48.991	0.613	48.975	0.613
sDiscrepancy or RE[a]	*1.022*	*1.079*	*1.280*	*1.058*	*1.253*	*1.058*
σ^2_y	58.653	6.994	58.220	6.854	57.790	6.790
sDiscrepancy or RE[a]	*-0.980*	*1.054*	*-1.044*	*1.013*	*-1.107*	*0.994*
$cov_{x,y}$	35.718	5.331	34.881	5.238	34.398	5.199
sDiscrepancy or RE[a]	*-0.469*	*1.041*	*-0.629*	*1.005*	*-0.722*	*0.990*
$\mu_{y\text{-}x}$	4.048	0.540	4.203	0.541	4.187	0.546
sDiscrepancy or RE[a]	*1.248*	*1.230*	*1.567*	*1.234*	*1.534*	*1.257*

Note. μ_y = mean posttest NW-9 score; σ^2_y = posttest NW-9 score variance; $cov_{x,y}$ = covariance between pretest and posttest NW-9 scores; $\mu_{y\text{-}x}$ = mean pre-post NW-9 score growth

[a] Standardized discrepancy (sDiscrepancy) is reported for parameter estimates, and relative efficiency (RE) is reported for standard errors. Standardized discrepancy quantifies the standard error difference between the parameter estimate obtained utilizing the missing data handling method and the complete data parameter estimate. Thus, standardized discrepancy values can be interpreted similarly to

z-scores, with values greater than ~|2| considered large and highlighted. Standardized discrepancy was small for all parameters across all methods. Relative efficiency quantifies the ratio between the squared standard errors obtained utilizing the missing data handling method and the squared standard errors obtained utilizing the complete dataset. Relative efficiency values can also be interpreted as the factor the sample size should be increased for a given missing data handling method to achieve the same standard errors as the complete dataset. For instance, the RE value for the listwise μ_y is 1.178, indicating that the listwise sample size should be increased by 17.8% to achieve the same μ_y standard error that was obtained using the complete dataset. Relative efficiency values greater than 1.2 were considered large and highlighted. No relative efficiency estimates indicated substantial standard error inflation. Only the standard error associated with the pre-post mean difference showed substantial inflation, and this inflation was fairly consistent across missing data methods. Interestingly, MI utilizing university and pretest auxiliary variables did not show substantial pre-post mean difference standard error inflation, but this result may be idiosyncratic of the 20 imputations in this condition.

Table 26

Comparison of NW-9 Results Across Different Missing Data Handling Techniques (50% Missingness)

	Complete		Listwise		MI (no aux)		MI (U. vars and pretest aux only)		MI (all aux)	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
μ_y	48.130	0.837	49.391	0.931	49.905	0.831	50.185	0.852	49.986	0.860
sDiscrepancy or RE[a]	---	---	*1.507*	*1.237*	*2.121*	*0.986*	*2.455*	*1.036*	*2.217*	*1.056*
σ^2_y	64.437	9.500	39.890	8.318	40.085	8.549	46.239	9.568	52.113	11.251
sDiscrepancy or RE[a]	---	---	*-2.584*	*0.767*	*-2.563*	*0.810*	*-1.916*	*1.014*	*-1.297*	*1.403*
$cov_{x,y}$	38.912	7.605	28.897	8.157	30.201	6.861	32.402	7.193	33.083	7.668
sDiscrepancy or RE[a]	---	---	*-1.317*	*1.150*	*-1.145*	*0.814*	*-0.856*	*0.895*	*-0.766*	*1.017*
$\mu_{y\text{-}x}$	3.554	0.705	5.739	0.908	5.328	0.822	5.609	0.815	5.410	0.814
sDiscrepancy or RE[a]	---	---	*3.099*	*1.659*	*2.516*	*1.359*	*2.915*	*1.336*	*2.633*	*1.333*

	FIML (no aux)		FIML (U. vars and pretest aux only)		FIML (all aux)	
	Est.	SE	Est.	SE	Est.	SE
μ_y	49.870	0.850	50.177	0.829	***	***
sDiscrepancy or RE[a]	*2.079*	*1.031*	*2.446*	*0.981*	***	***
σ^2_y	40.776	8.298	43.555	8.622	***	***
sDiscrepancy or RE[a]	*-2.491*	*0.763*	*-2.198*	*0.824*	***	***
$cov_{x,y}$	30.606	7.367	32.328	7.147	***	***
sDiscrepancy or RE[a]	*-1.092*	*0.938*	*-0.866*	*0.883*	***	***
$\mu_{y\text{-}x}$	5.294	0.836	5.601	0.791	***	***
sDiscrepancy or RE[a]	*2.468*	*1.406*	*2.904*	*1.259*	***	***

Note. μ_y = mean posttest NW-9 score; σ^2_y = posttest NW-9 score variance; $cov_{x,y}$ = covariance between pretest and posttest NW-9 scores; $\mu_{y\text{-}x}$ = mean pre-post NW-9 score growth. FIML estimation utilizing all auxiliary variables was not able to converge on a solution after 10,000 replications.

[a] Standardized discrepancy (sDiscrepancy) is reported for parameter estimates, and relative efficiency (RE) is reported for standard errors. Standardized discrepancy quantifies the standard error difference between the parameter estimate obtained utilizing the missing

data handling method and the complete data parameter estimate. Thus, standardized discrepancy values can be interpreted similarly to *z*-scores, with values greater than ~|2| considered large and highlighted. Across all conditions, pre-post mean change estimates were substantially larger when utilizing a missing data treatment technique than when analyzing the complete data. Posttest variance estimates were substantially smaller when utilizing a missing data treatment technique than those obtained when analyzing the complete data, but this bias was reduced when more auxiliary variables were used. Interestingly, all missing data techniques resulted in a posttest mean estimate larger than that obtained by analyzing the complete data, with the exception of listwise deletion. Relative efficiency quantifies the ratio between the squared standard errors obtained utilizing the missing data handling method and the squared standard errors obtained utilizing the complete dataset. Relative efficiency values can also be interpreted as the factor the sample size should be increased for a given missing data handling method to achieve the same standard errors as the complete dataset. For instance, the RE value for the listwise μ_y is 1.237, indicating that the listwise sample size should be increased by 23.7% to achieve the same μ_y standard error that was obtained using the complete dataset. Relative efficiency values greater than 1.2 were considered large and highlighted. Standard error inflation was most problematic for pre-post mean change estimate standard errors across missing data techniques

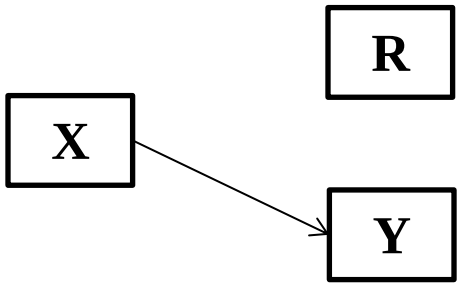


Figure 1a. MCAR model. Missingness (*R*) is unrelated to both other variables in the dataset (*X*) and to the variable with missingness (*Y*).

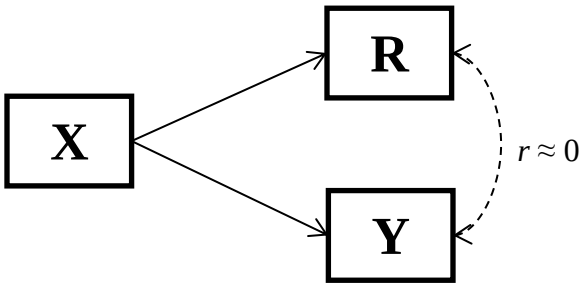


Figure 1b. MAR model. Missingness (*R*) is unrelated to the variable with missingness (*Y*) *after controlling for* the other variables in the dataset (*X*).

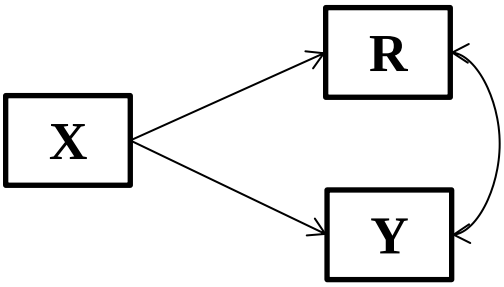


Figure 1c. MNAR model. Missingness (*R*) is related to the variable with missingness (*Y*) even after controlling for the other variables in the dataset (*X*).

DATASET 1

	PRE	POST	MAKEUP
Student 1	X	X	
Student 2	X	X	
Student 3	X	X	
Student 4	X	X	
Student 5	X		X
Student 6	X		X

DATASET 2

	PRE	POST	MAKEUP
Student 1	X	X	
Student 2	X	X	
Student 3	X	X	
Student 4	X	X	
Student 5	X		X
Student 6	X		X

DATASET 3

	PRE	POST	MAKEUP
Student 1	X	X	
Student 2	X	X	
Student 3	X	X	
Student 4	X	X	
Student 5	X		X
Student 6	X		X

Figure 2. Different pre-post datasets. X's denote present data. Dataset 1 involves listwise deleting Students 5 and 6, whose posttest data was obtained during a makeup testing session. Dataset 2 involves using the complete dataset, including both standard posttest and makeup posttest data. Dataset 3 involves treating makeup posttest data as missing, and utilizing MI or FIML missing data techniques to handle the missing posttest scores for Students 5 and 6.

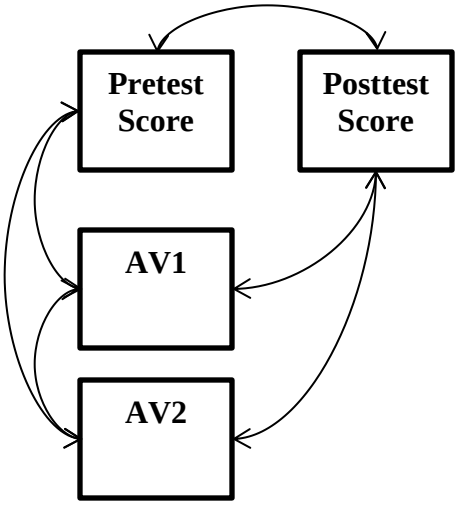


Figure 3. Incorporating auxiliary variables into FIML analysis of pretest and posttest scores. AV = Auxiliary variable. Auxiliary variables are allowed to correlate with each other, as well as pretest and posttest scores. Although only two auxiliary variables are shown in the diagram, additional auxiliary variables (see Table 4) will be utilized.

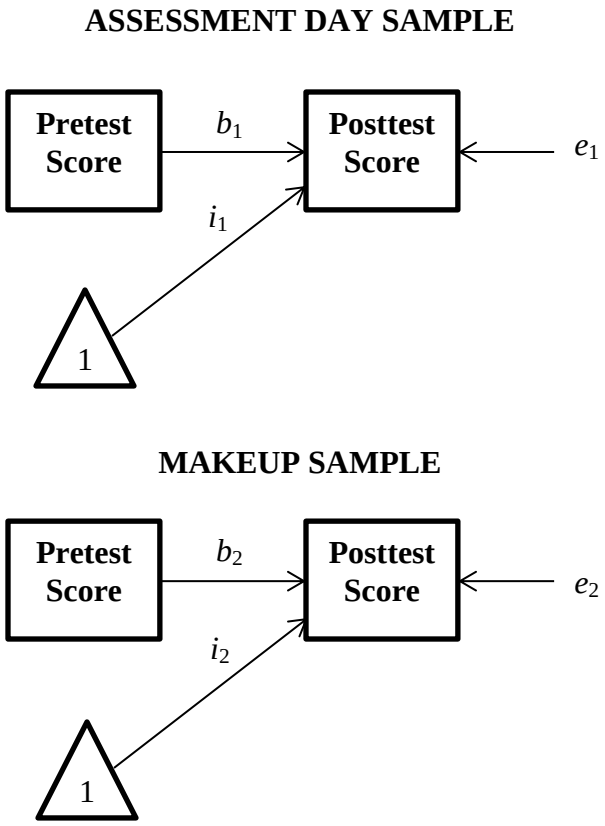


Figure 4. Multiple-group analysis to examine potential random responding by posttest makeup students. The fit of models constraining intercepts (i_1 and i_2), slopes (b_1 and b_2), residual variances (e_1 and e_2), or all three to be equivalent across samples were assessed. If the model with equivalent intercepts, slopes, and residual variances across groups was associated with no significant misfit, this lack of misfit would indicate the relationship between the two constructs does not vary across groups. If the makeup sample has a diminished intercept, diminished pre-post slope, and/or increased residual variance compared to the Assessment Day sample, these differences may indicate that makeup students are responding randomly at posttest.

Appendix A

Sample Syntax for Listwise Deletion and Complete Data Conditions

```
DATA: file = mapLIST.csv;

!Listing out variables, but only using pretest and posttest

VARIABLE:
names = id attend sp09map FA07map gender sp09age sat1math
sat1verb GPA credhrs fa07pap fa07mav fa07pav fa07wav
FA07ope FA07con FA07ext FA07agr FA07neu FA07mair fa07effort
fa07import sp09pap sp09mav sp09pav sp09wav sp09ope sp09con
sp09ext sp09agr sp09neu sp09eff sp09imp;

usevariables = SP09map FA07map;

!Using the maximum-likelihood estimator

ANALYSIS:
estimator = ml;

MODEL:
!Pretest mean
[FA07MAP] (premean);
!Posttest mean
[SP09MAP] (postmean);
!Pretest and Posttest variances
FA07MAP SP09MAP;
!Pretest-Posttest covariance
FA07MAP with SP09MAP;

!Estimating pre-post mean difference
MODEL CONSTRAINT:
new(meandiff);
meandiff = postmean-premean;

!Output will give sample statistics, patterns of
!missingness and standardized solution
OUTPUT:
sampstat patterns stdyx;
```

Note. Exclamation marks (!) denote comments. Listwise and complete datasets will differ only in the dataset being read into MPlus.

Appendix B

Sample Syntax for MI Imputation Phase Excluding Auxiliary Variables

```
DATA: file = mapMISS.csv;

!Listing out variables, and but only using posttest and
!pretest MAP scores

VARIABLE:
names = id attend sp09map FA07map gender sp09age sat1math
sat1verb GPA credhrs fa07pap fa07mav fa07pav fa07wav
FA07ope FA07con FA07ext FA07agr FA07neu FA07mair fa07effort
fa07import sp09pap sp09mav sp09pav sp09wav sp09ope sp09con
sp09ext sp09agr sp09neu sp09eff sp09imp;

usevariables = sp09map FA07map;

!Missing variable code
missing = all (-9);

!Providing Bayes seed and convergence criteria
!for imputation

ANALYSIS:
Type = basic;
Bseed = 467484;
Bconvergence = .01;

!Imputing posttest scores, 20 datasets, extracting
!every 5000th imputation

DATA IMPUTATION:
Impute = sp09map;
Ndatasets = 20;
Save = MAPMInoaux*.dat;
Thin = 5000;

!Tech8 monitors imputation convergence process

OUTPUT:
Tech8;
```

Note. Exclamation marks (!) denote comments.

Appendix C

Sample Syntax for MI Imputation Phase Including University Database and Pretest Auxiliary Variables

```
DATA: file = mapMISS.csv;

!Listing out all variables – all auxiliary variables being
!used in the imputation process – note that posttest
!auxiliary variables are excluded.
VARIABLE:
names = id attend sp09map FA07map gender sp09age sat1math
sat1verb GPA credhrs fa07pap fa07mav fa07pav fa07wav
FA07ope FA07con FA07ext FA07agr FA07neu FA07mair fa07effort
fa07import sp09pap sp09mav sp09pav sp09wav sp09ope sp09con
sp09ext sp09agr sp09neu sp09eff sp09imp;
usevariables = sp09map FA07map gender sp09age sat1math
sat1verb GPA credhrs fa07pap fa07mav fa07pav fa07wav
FA07ope FA07con FA07ext FA07agr FA07neu FA07mair fa07effort
fa07import;

!Missing variable code
missing = all (-9);

!Providing Bayes seed and convergence criteria
!for imputation
ANALYSIS:
Type = basic;
Bseed = 186746;
Bconvergence = .01;

!Imputing posttest scores, as well as auxiliary variables
!with missing values, 20 datasets, extracting every 5000th
!imputation
DATA IMPUTATION:
Impute = sp09map gender sp09age sat1math sat1verb GPA
credhrs fa07pap fa07mav fa07pav fa07wav FA07ope FA07con
FA07ext FA07agr FA07neu FA07mair fa07effort fa07import;
Ndatasets = 20;
Save = MAPMIpreaux*.dat;
Thin = 5000;

!Tech8 monitors imputation convergence process
OUTPUT: Tech8;
```

Note. Exclamation marks (!) denote comments. Although all variables are used in the imputation model in this example, the selection of auxiliary variables will be dependent on screening for relationships with missingness and posttest scores.

Appendix D

Sample Syntax for MI Imputation Phase Including All Auxiliary Variables

```
DATA: file = mapMISS.csv;

!Listing out all variables – all auxiliary variables being
!used in the imputation process.

VARIABLE:
names = id attend sp09map FA07map gender sp09age sat1math
sat1verb GPA credhrs fa07pap fa07mav fa07pav fa07wav
FA07ope FA07con FA07ext FA07agr FA07neu FA07mair fa07effort
fa07import sp09pap sp09mav sp09pav sp09wav sp09ope sp09con
sp09ext sp09agr sp09neu sp09eff sp09imp;
usevariables = sp09map FA07map gender sp09age sat1math
sat1verb GPA credhrs fa07pap fa07mav fa07pav fa07wav
FA07ope FA07con FA07ext FA07agr FA07neu FA07mair fa07effort
fa07import sp09pap sp09mav sp09pav sp09wav sp09ope sp09con
sp09ext sp09agr sp09neu sp09eff sp09imp;

!Missing variable code
missing = all (-9);

!Providing Bayes seed and convergence criteria
!for imputation
ANALYSIS:
Type = basic;
Bseed = 973732;
Bconvergence = .01;

!Imputing posttest scores, as well as auxiliary variables
!with missing values, 20 datasets, extracting every 5000th
!imputation
DATA IMPUTATION:
Impute = sp09map gender sp09age sat1math sat1verb GPA
credhrs fa07pap fa07mav fa07pav fa07wav FA07ope FA07con
FA07ext FA07agr FA07neu FA07mair fa07effort fa07import
sp09pap sp09mav sp09pav sp09wav sp09ope sp09con sp09ext
sp09agr sp09neu sp09eff sp09imp;
Ndatasets = 20;
Save = MAPMIallaux*.dat;
Thin = 5000;

!Tech8 monitors imputation convergence process
OUTPUT:
Tech8;
```

Note. Exclamation marks (!) denote comments. Although all variables are used in the imputation model in this example, the selection of auxiliary variables will be dependent on screening for relationships with missingness and posttest scores.

Appendix E

Sample Syntax for MI Analysis Phase

```
DATA: file = MAPMInoauxlist.dat;

!Indicates that the data file is a list of multiple imputed
!datasets
Type = imputation;

!Only need to use pretest and posttest scores in the
!analysis model
VARIABLE:
names = FA07MAP SP09MAP;
usevariables = FA07MAP SP09MAP;

!Using the maximum-likelihood estimator
ANALYSIS:
estimator = ml;

MODEL:
!Pretest mean
[FA07MAP] (premean);
!Posttest mean
[SP09MAP] (postmean);
!Pretest and Posttest variances
FA07MAP SP09MAP;
!Pretest-Posttest covariance
FA07MAP with SP09MAP;

!Estimating pre-post mean difference
MODEL CONSTRAINT:
new(meandiff);
meandiff = postmean-premean;

!Output give sample statistics, patterns of missingness
!and standardized solution
OUTPUT:
sampstat patterns stdyx;
```

Note. Exclamation marks (!) denote comments. This syntax analyzes imputed data associated with Appendix B, excluding auxiliary variables. The syntax analyzing imputed data associated with Appendix C would replace the data file with "MAPMIpreauxlist.dat" and the variable list with those imputed in Appendix C, and the syntax analyzing imputed data associated with Appendix D would replace the data file with "MAPMIallauxlist.dat" and the variable list with those imputed in Appendix D.

Appendix F

Sample Syntax for FIML Analysis Excluding Auxiliary Variables

```
DATA: file = MAPMISS.csv;

!Listing out variables, but only using pretest and posttest

VARIABLE:
names = id attend sp09map FA07map gender sp09age sat1math
sat1verb GPA credhrs fa07pap fa07mav fa07pav fa07wav
FA07ope FA07con FA07ext FA07agr FA07neu FA07mair fa07effort
fa07import sp09pap sp09mav sp09pav sp09wav sp09ope sp09con
sp09ext sp09agr sp09neu sp09eff sp09imp;
usevariables = FA07MAP SP09MAP;

!Missing variable code
missing = all (-9);

!Using the maximum-likelihood estimator
ANALYSIS:
estimator = ml;

MODEL:
!Pretest mean
[FA07MAP] (premean);
!Posttest mean
[SP09MAP] (postmean);
!Pretest and Posttest variances
FA07MAP SP09MAP;
!Pretest-Posttest covariance
FA07MAP with SP09MAP;

!Estimating pre-post mean difference
MODEL CONSTRAINT:
new(meandiff);
meandiff = postmean-premean;

!Output give sample statistics, patterns of missingness
!and standardized solution
OUTPUT:
sampstat patterns stdyx;
```

Note. Exclamation marks (!) denote comments.

Appendix G

Sample Syntax for FIML Analysis Including University Database and Pretest Auxiliary Variables

```
DATA: file = MAPMISS.csv;

VARIABLE:
names = id attend sp09map FA07map gender sp09age sat1math
sat1verb GPA credhrs fa07pap fa07mav fa07pav fa07wav
FA07ope FA07con FA07ext FA07agr FA07neu FA07mair fa07effort
fa07import sp09pap sp09mav sp09pav sp09wav sp09ope sp09con
sp09ext sp09agr sp09neu sp09eff sp09imp;
usevariables = sp09map FA07map gender sp09age sat1math
sat1verb GPA credhrs fa07pap fa07mav fa07pav fa07wav
FA07ope FA07con FA07ext FA07agr FA07neu FA07mair fa07effort
fa07import;

!All variables being used as auxiliary variables
auxiliary = (m) gender sp09age sat1math sat1verb GPA
credhrs fa07pap fa07mav fa07pav fa07wav FA07ope FA07con
FA07ext FA07agr FA07neu FA07mair fa07effort fa07import;

!Missing variable code
missing = all (-9);

!Using the maximum-likelihood estimator
ANALYSIS: estimator = ml;

MODEL:
!Pretest mean
[FA07MAP] (premean);
!Posttest mean
[SP09MAP] (postmean);
!Pretest and Posttest variances
FA07MAP SP09MAP;
!Pretest-Posttest covariance
FA07MAP with SP09MAP;

!Estimating pre-post mean difference
MODEL CONSTRAINT:
new(meandiff);
meandiff = postmean-premean;

!Sample statistics, missingness patterns, and standardized
!solution
```

```
OUTPUT:
sampstat patterns stdyx;
```

Note. Exclamation marks (!) denote comments.

Appendix H

Sample Syntax for FIML Analysis Including All Auxiliary Variables

```
DATA: file = MAPMISS.csv;

VARIABLE:
names = id attend sp09map FA07map gender sp09age sat1math
sat1verb GPA credhrs fa07pap fa07mav fa07pav fa07wav
FA07ope FA07con FA07ext FA07agr FA07neu FA07mair fa07effort
fa07import sp09pap sp09mav sp09pav sp09wav sp09ope sp09con
sp09ext sp09agr sp09neu sp09eff sp09imp;
usevariables = sp09map FA07map gender sp09age sat1math
sat1verb GPA credhrs fa07pap fa07mav fa07pav fa07wav
FA07ope FA07con FA07ext FA07agr FA07neu FA07mair fa07effort
fa07import sp09pap sp09mav sp09pav sp09wav sp09ope sp09con
sp09ext sp09agr sp09neu sp09eff sp09imp;


!All variables being used as auxiliary variables
auxiliary = (m) gender sp09age sat1math sat1verb GPA
credhrs fa07pap fa07mav fa07pav fa07wav FA07ope FA07con
FA07ext FA07agr FA07neu FA07mair fa07effort fa07import
sp09pap sp09mav sp09pav sp09wav sp09ope sp09con sp09ext
sp09agr sp09neu sp09eff sp09imp;


!Missing variable code
missing = all (-9);

!Using the maximum-likelihood estimator
ANALYSIS: estimator = ml;

MODEL:
!Pretest mean
[FA07MAP] (premean);
!Posttest mean
[SP09MAP] (postmean);
!Pretest and Posttest variances
FA07MAP SP09MAP;
!Pretest-Posttest covariance
FA07MAP with SP09MAP;

!Estimating pre-post mean difference
MODEL CONSTRAINT:
new(meandiff);
meandiff = postmean-premean;
```

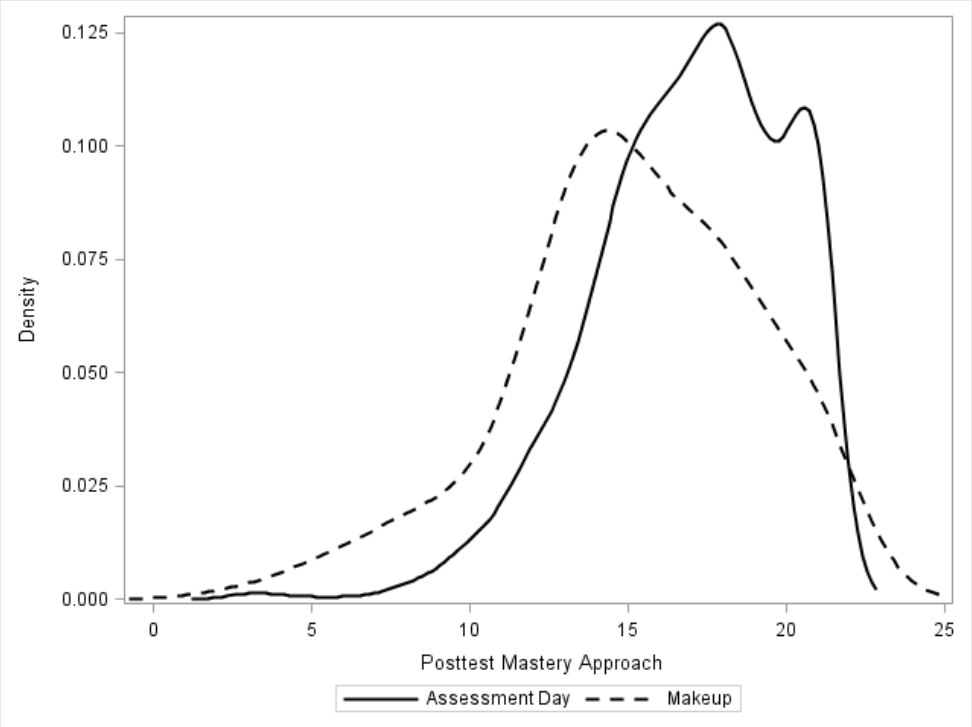
```
!Sample statistics, missingness patterns, and standardized
!solution
OUTPUT:
sampstat patterns stdyx;
```

Note. Exclamation marks (!) denote comments.

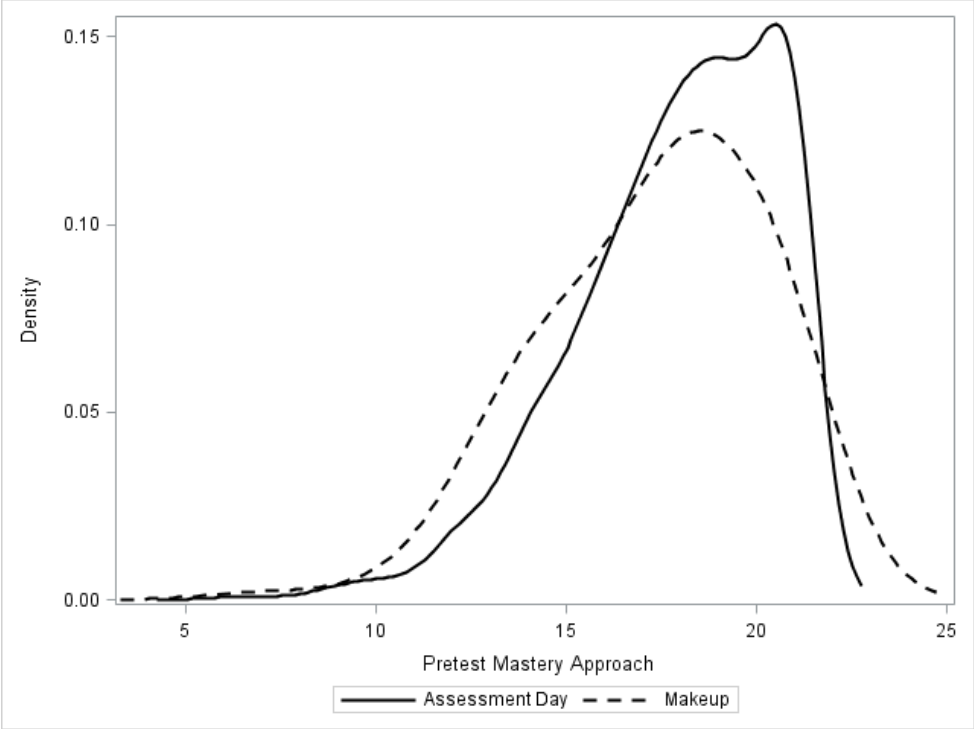
Appendix I

Histograms Comparing Assessment Day and Makeup Variable Distributions – Noncognitive Sample

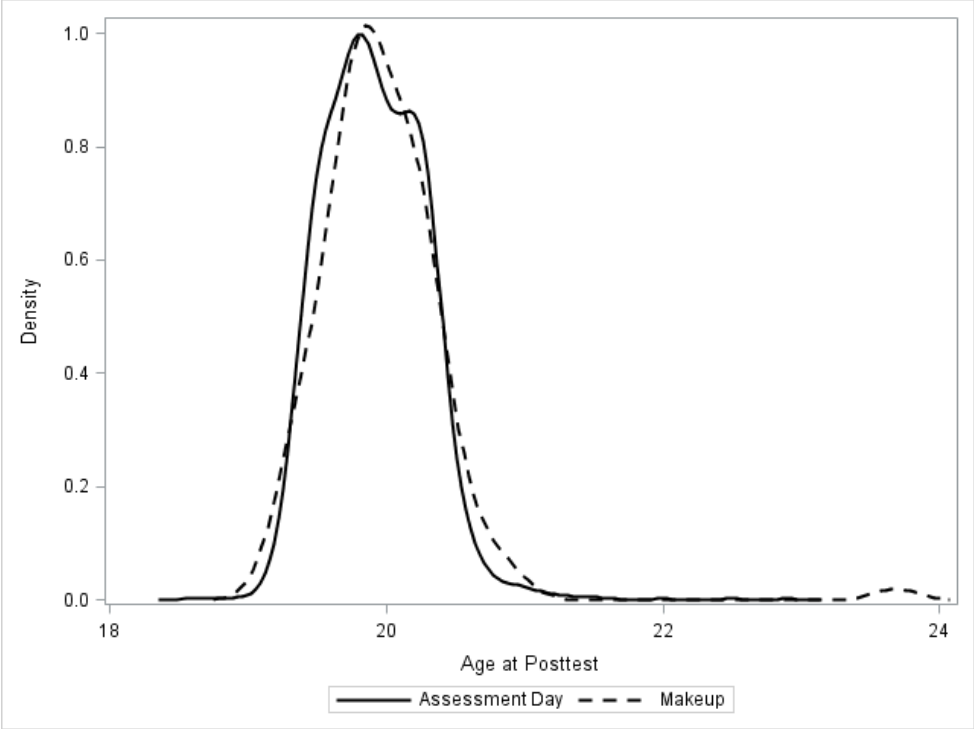
Posttest Mastery Approach



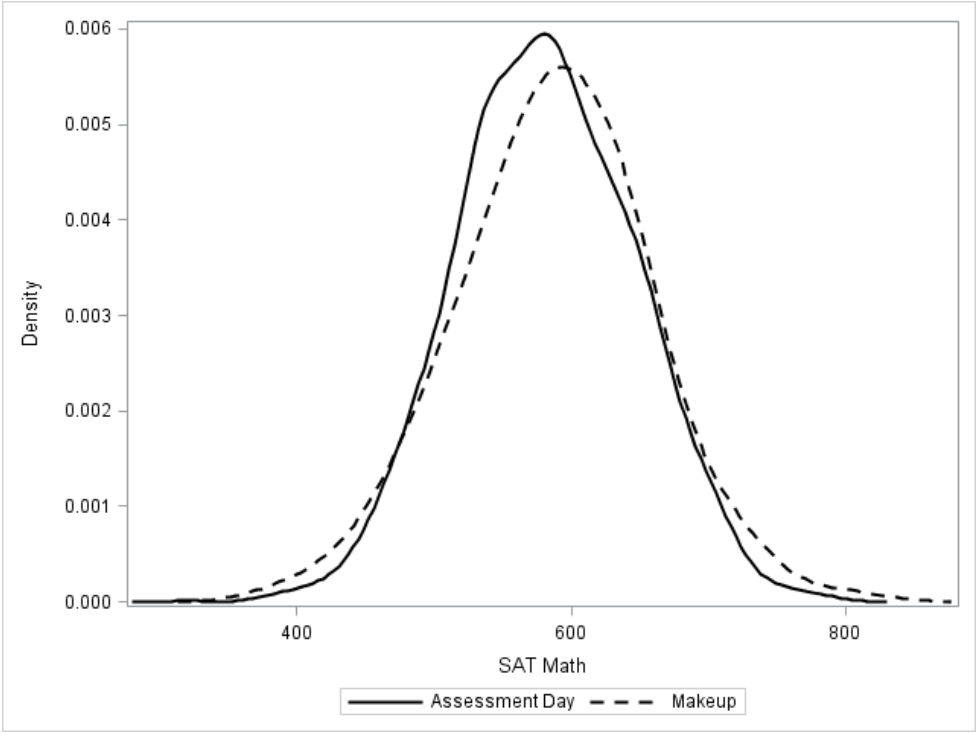
Pretest Mastery Approach



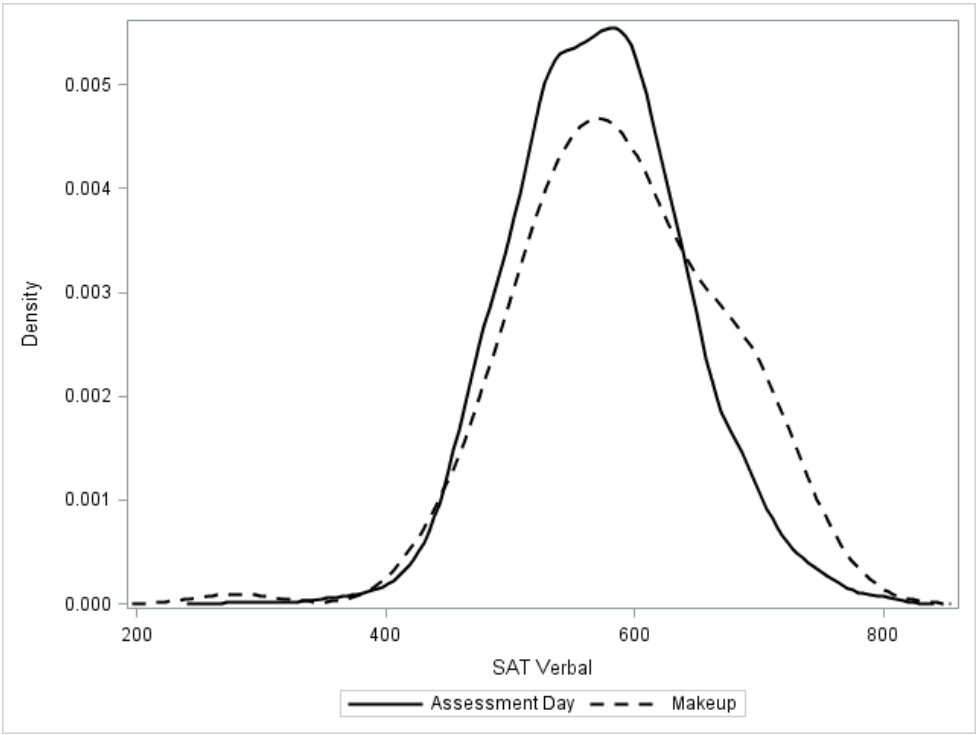
Age



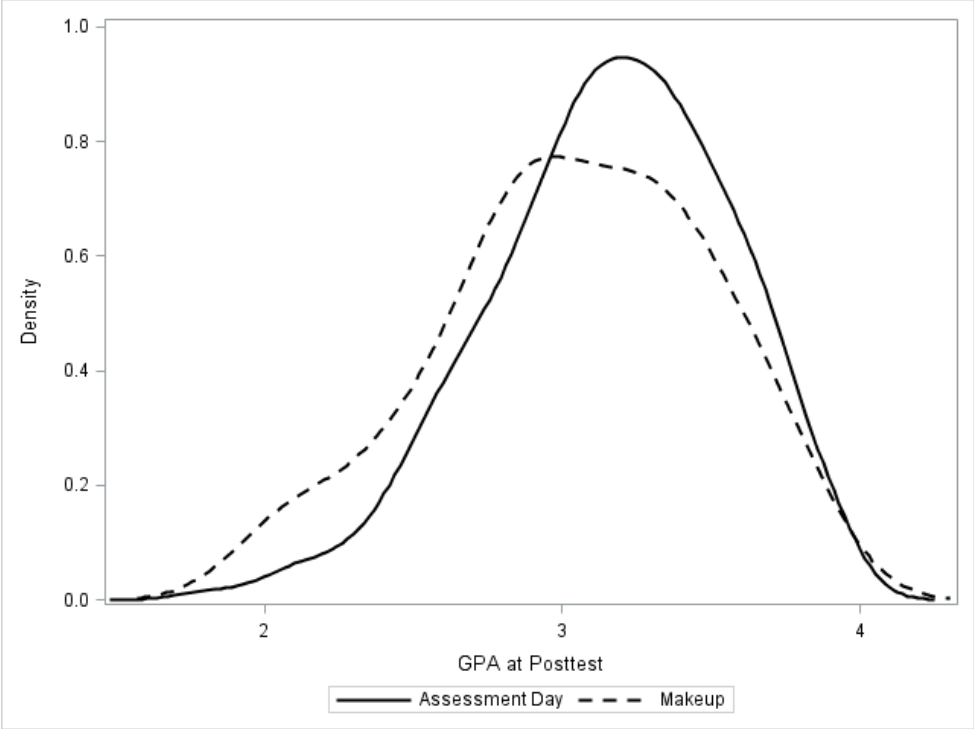
SAT Math



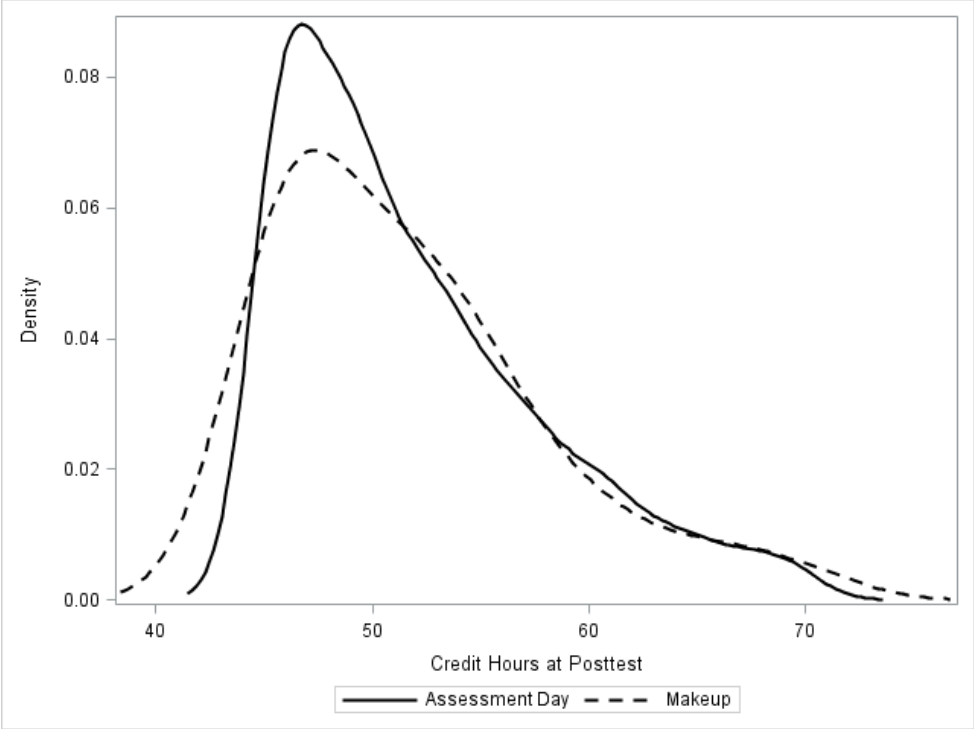
SAT Verbal



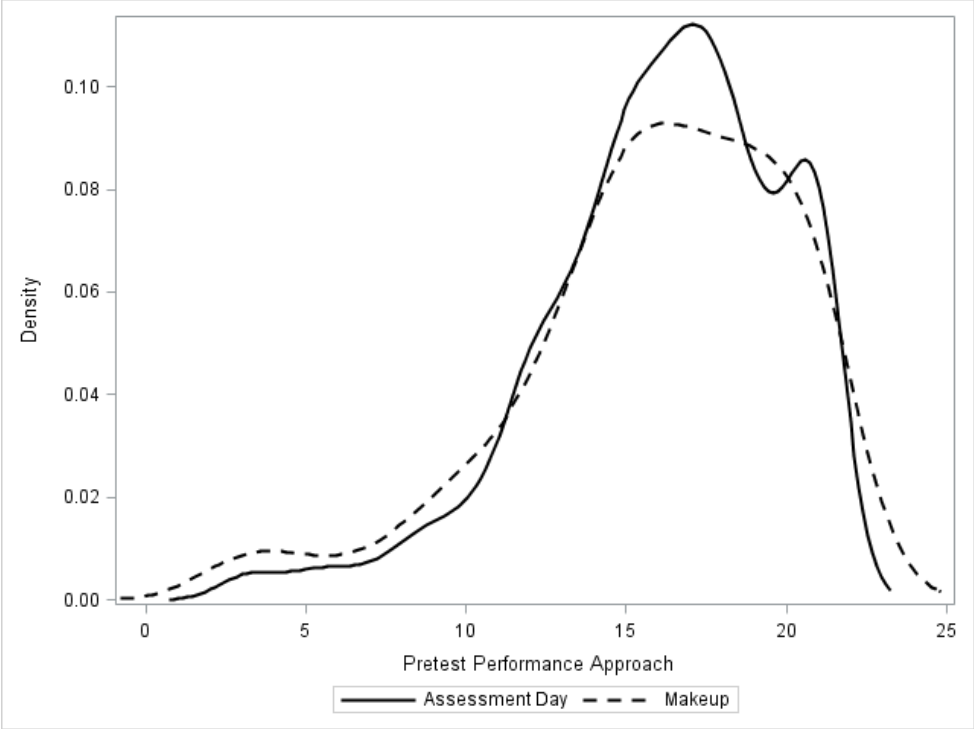
GPA at Posttest



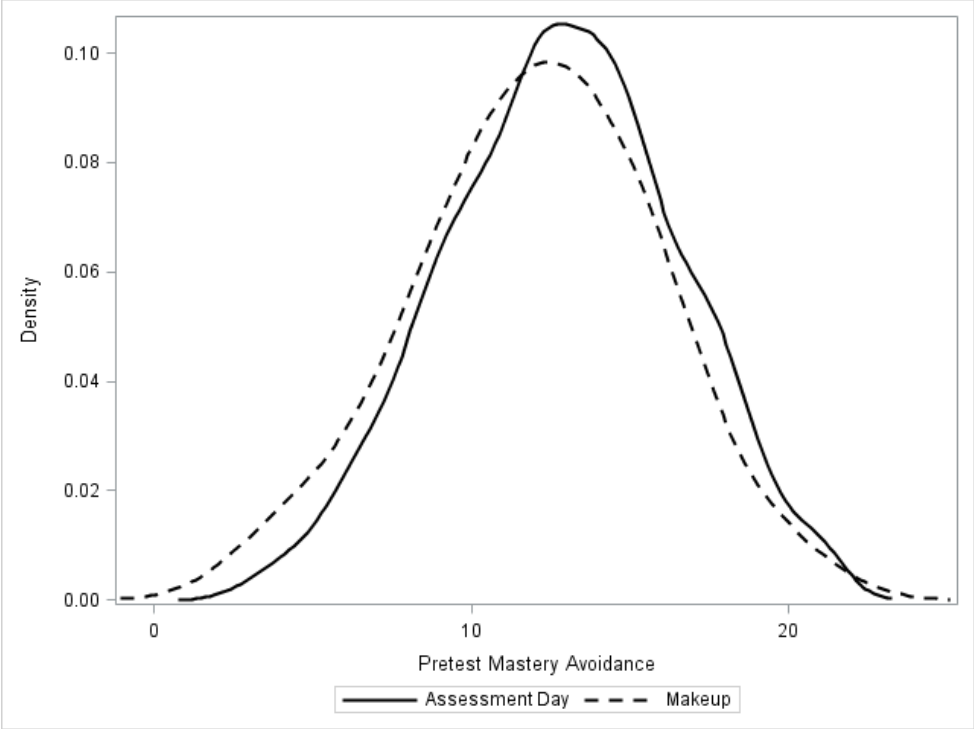
Credit Hours at Posttest



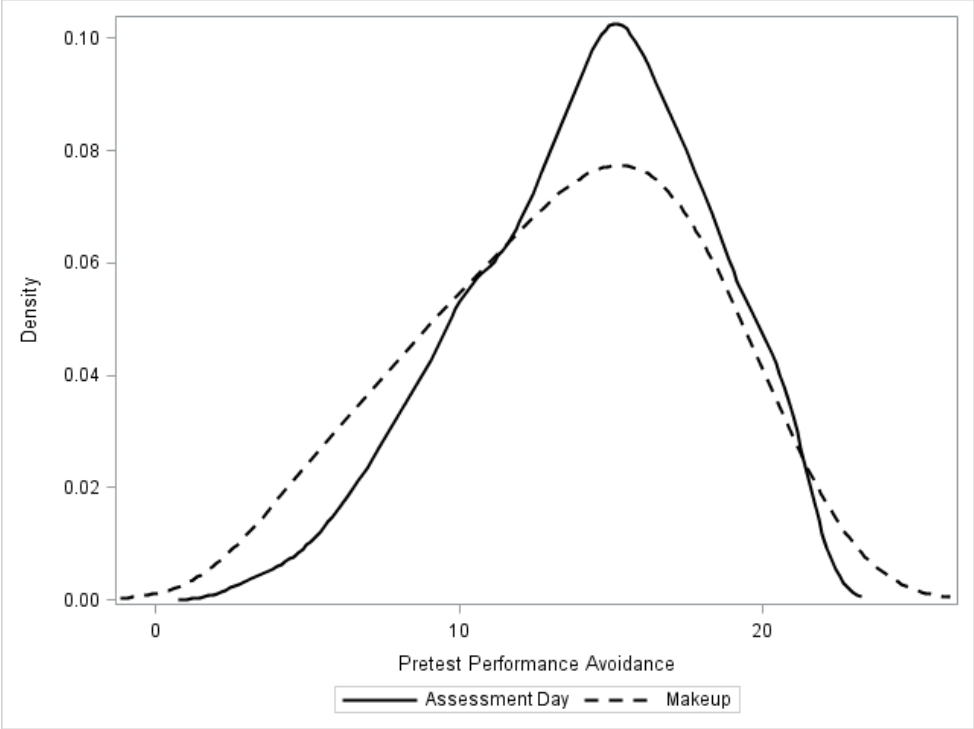
Pretest Performance Approach



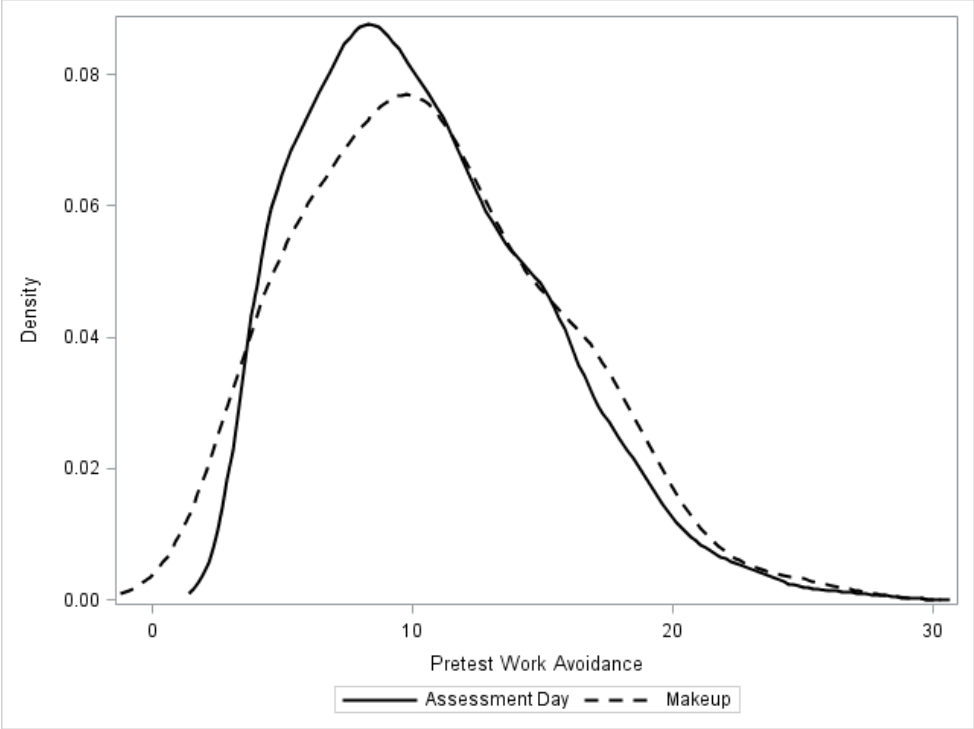
Pretest Mastery Avoidance



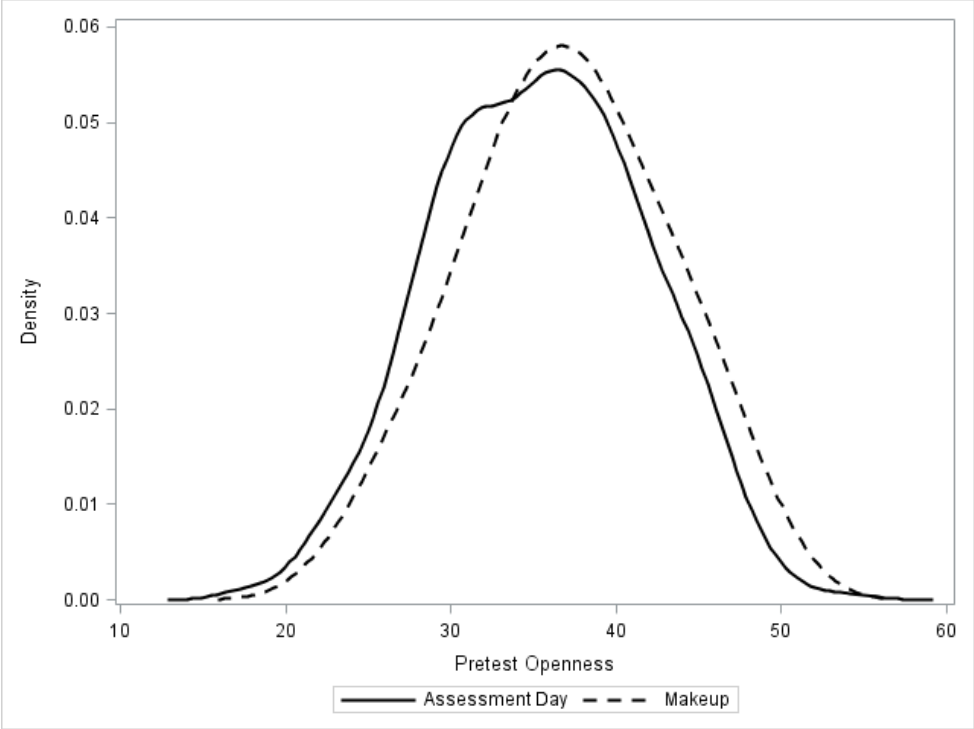
Pretest Performance Avoidance



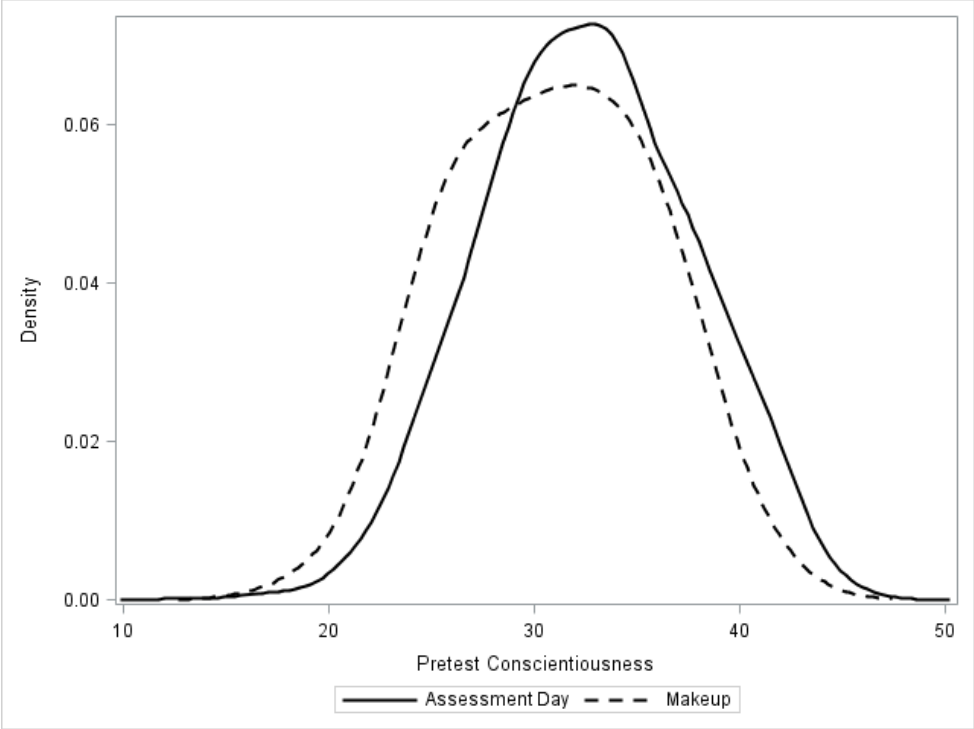
Pretest Work Avoidance



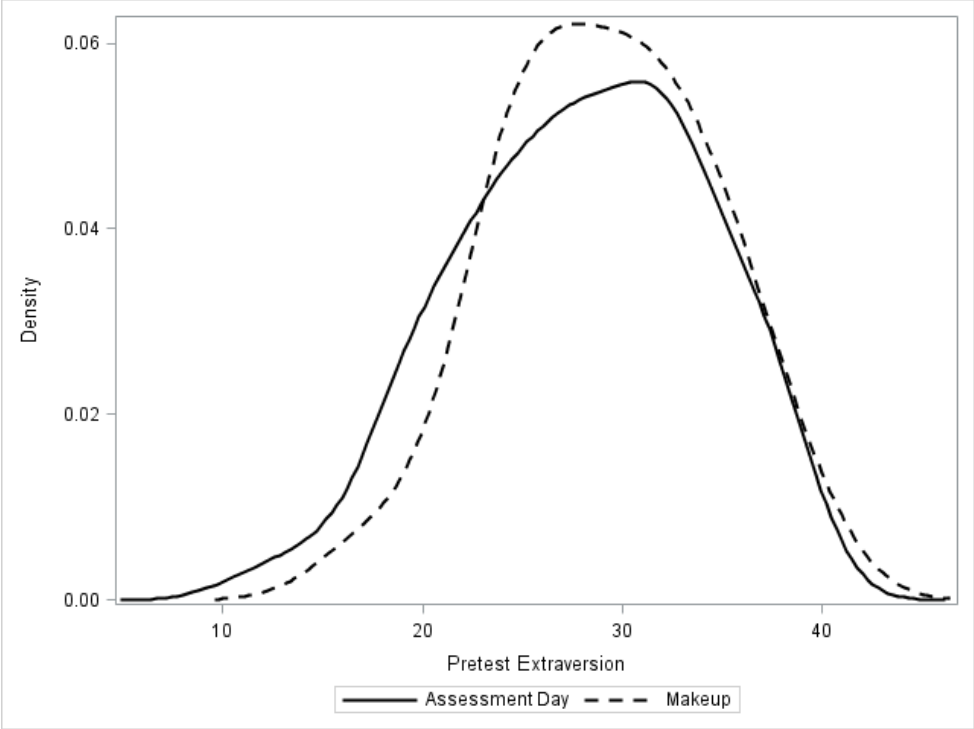
Pretest Openness



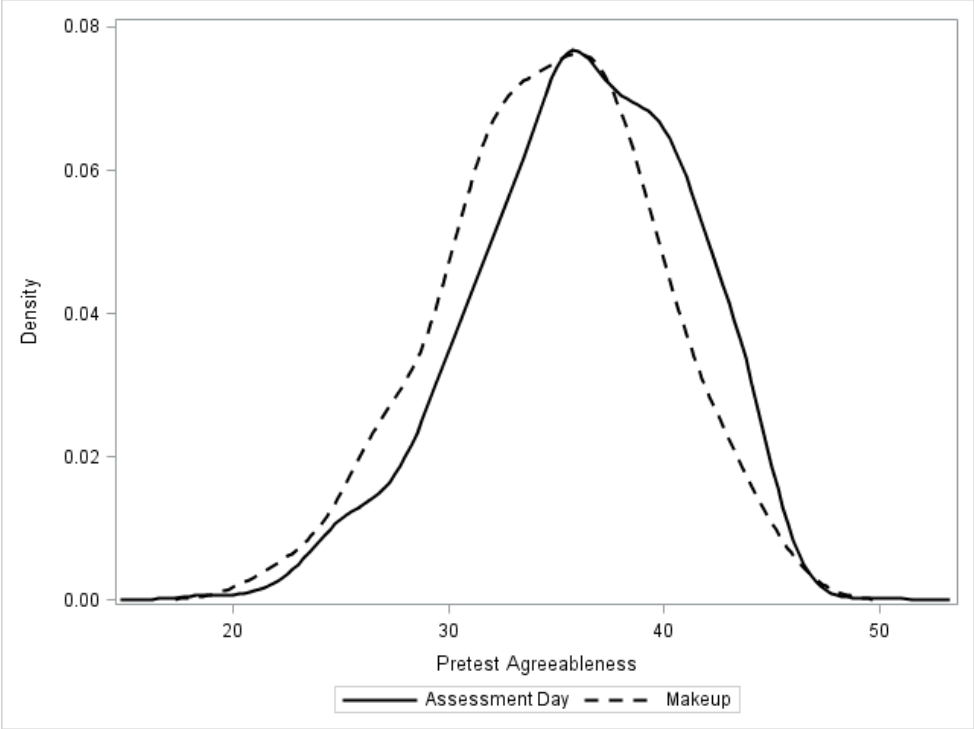
Pretest Conscientiousness



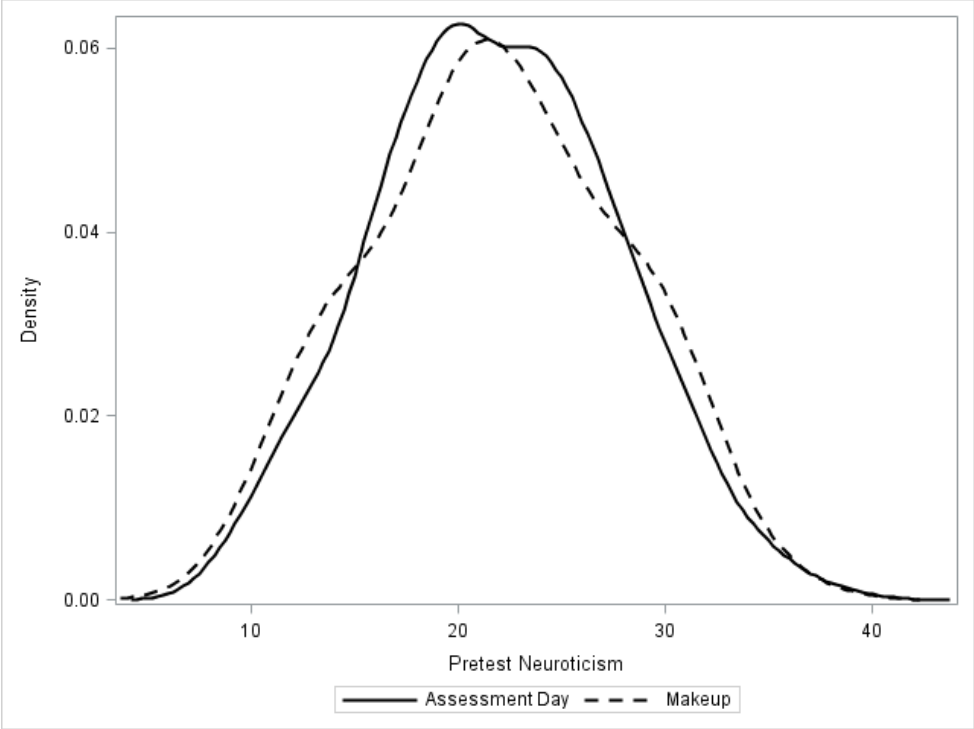
Pretest Extraversion



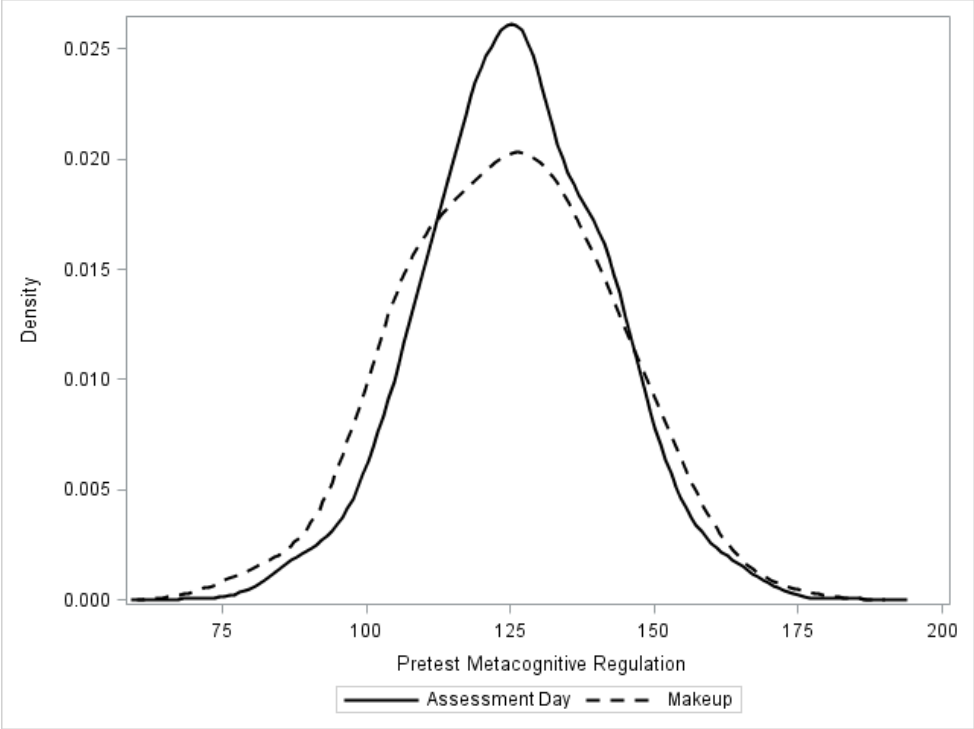
Pretest Agreeableness



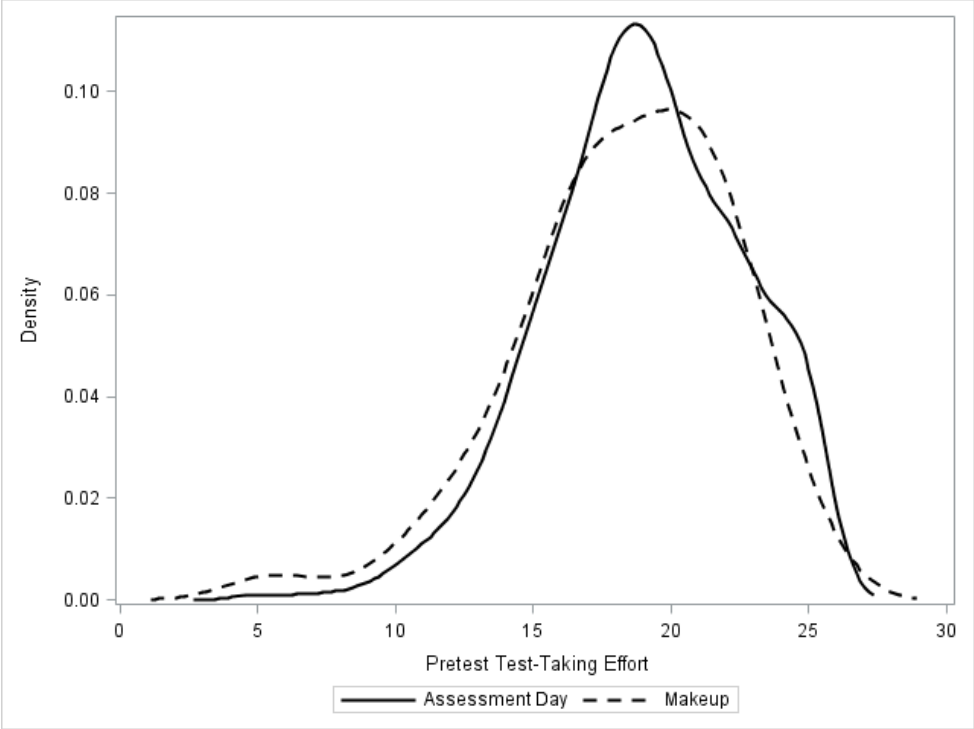
Pretest Neuroticism



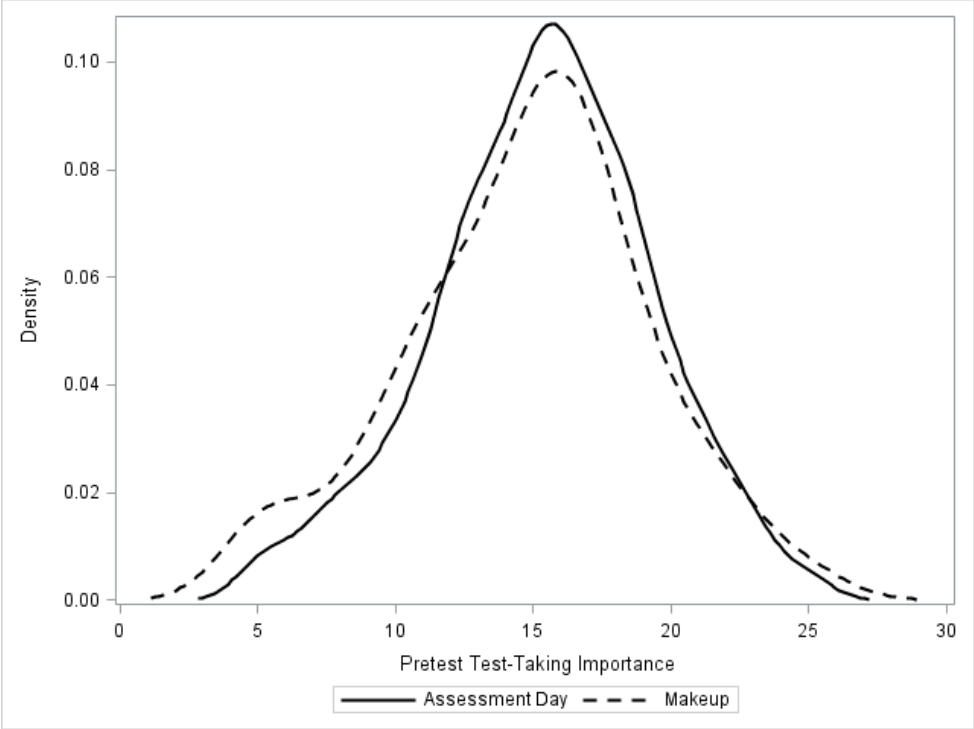
Pretest Metacognitive Regulation



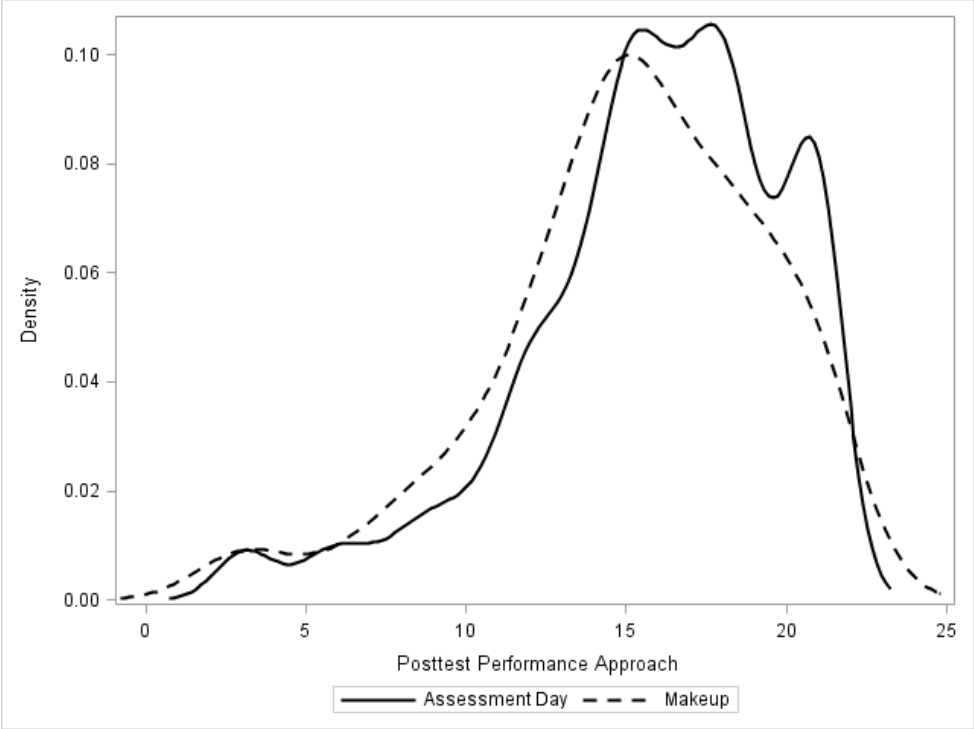
Pretest Test-taking Effort



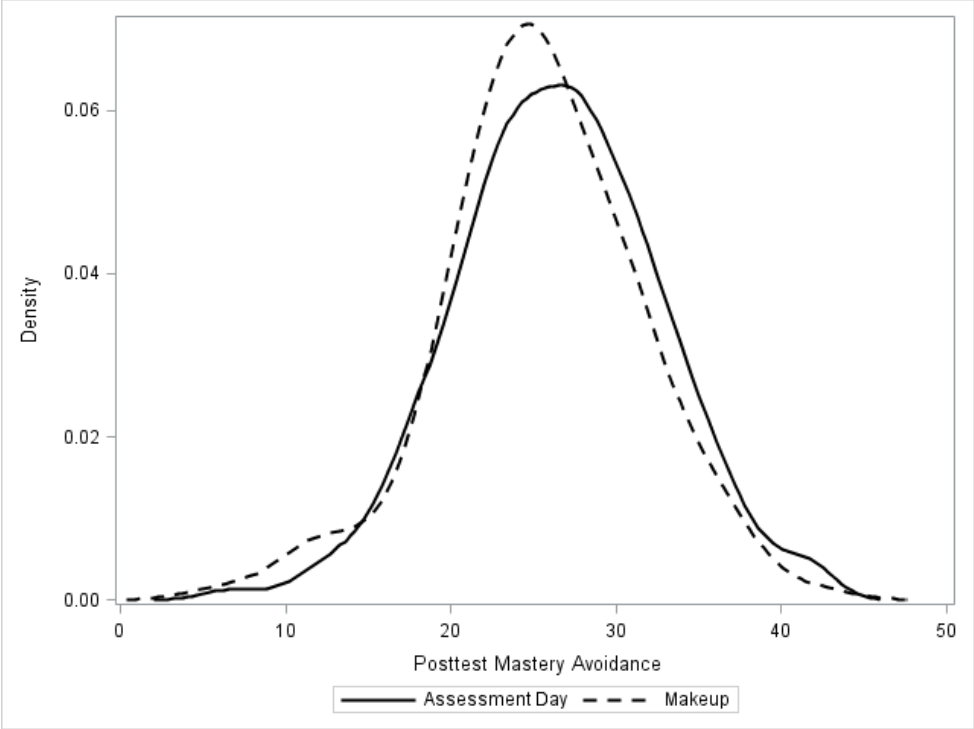
Pretest Test-taking Importance



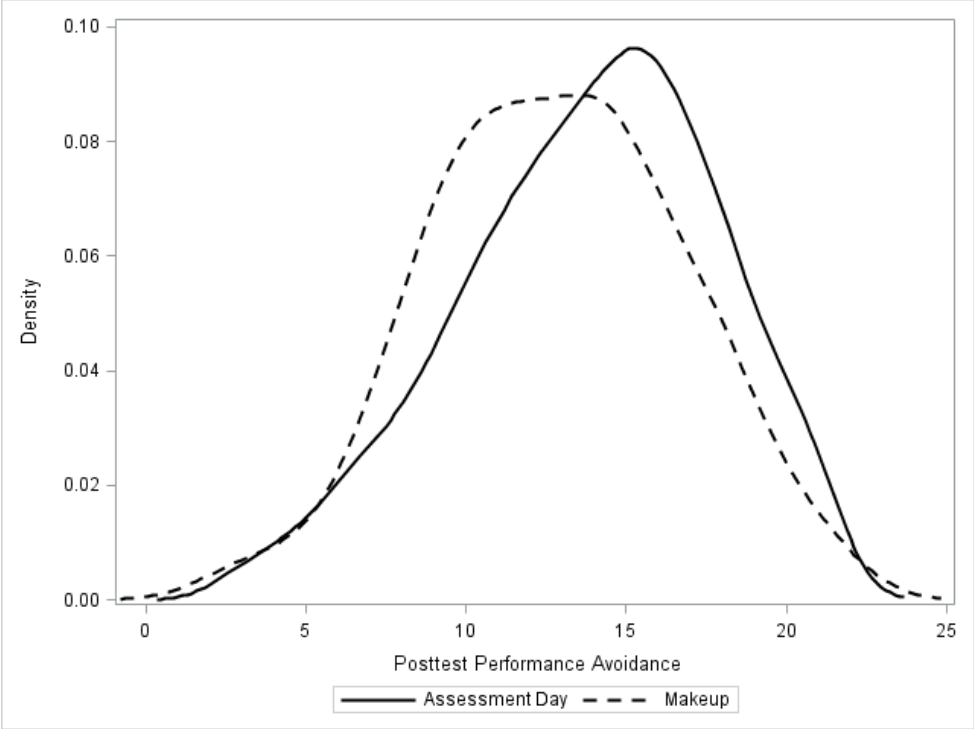
Posttest Performance Approach



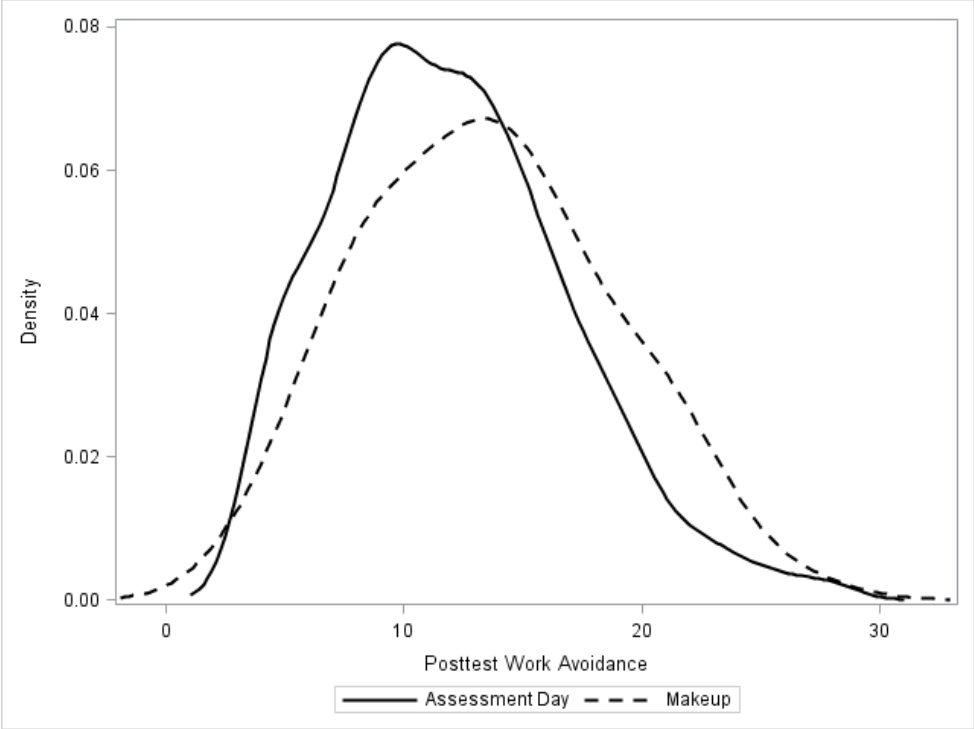
Posttest Mastery Avoidance



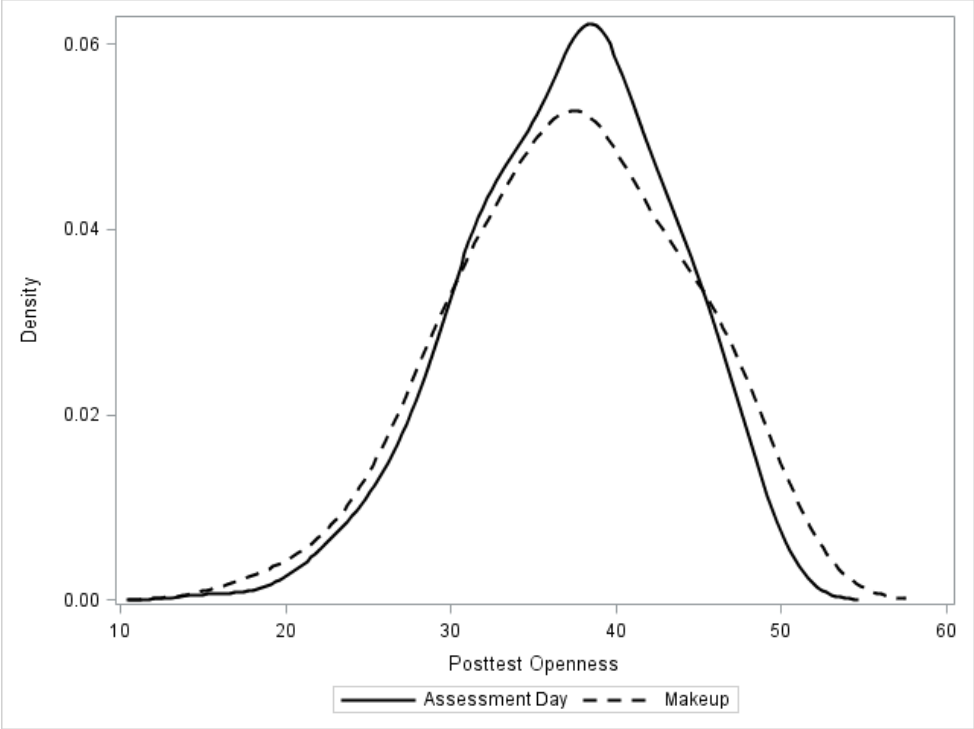
Posttest Performance Avoidance



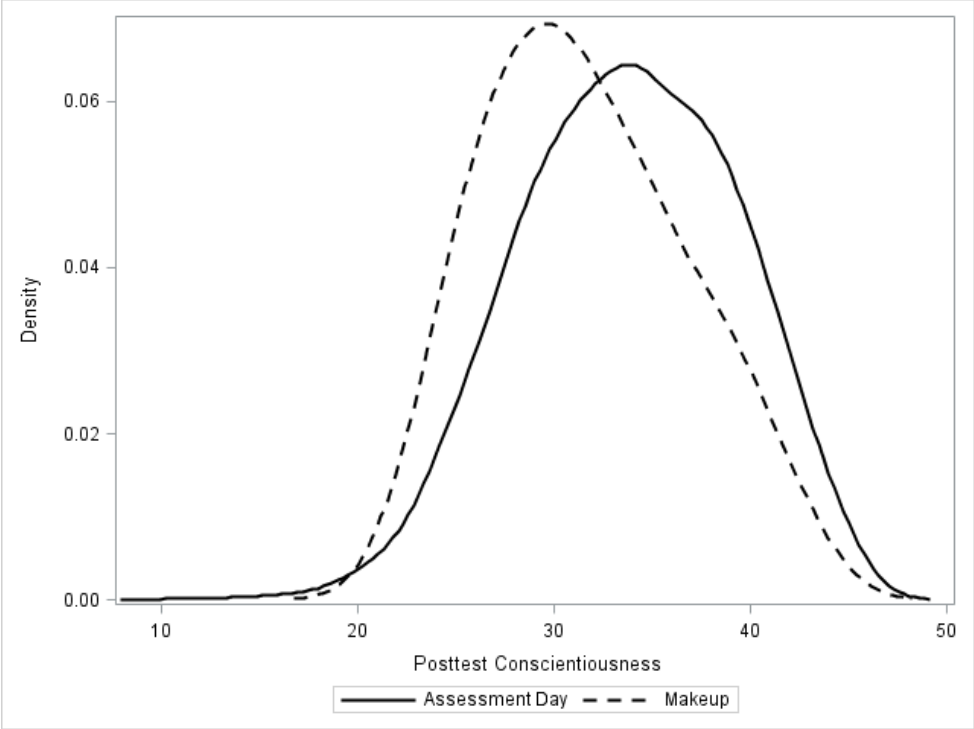
Posttest Work Avoidance



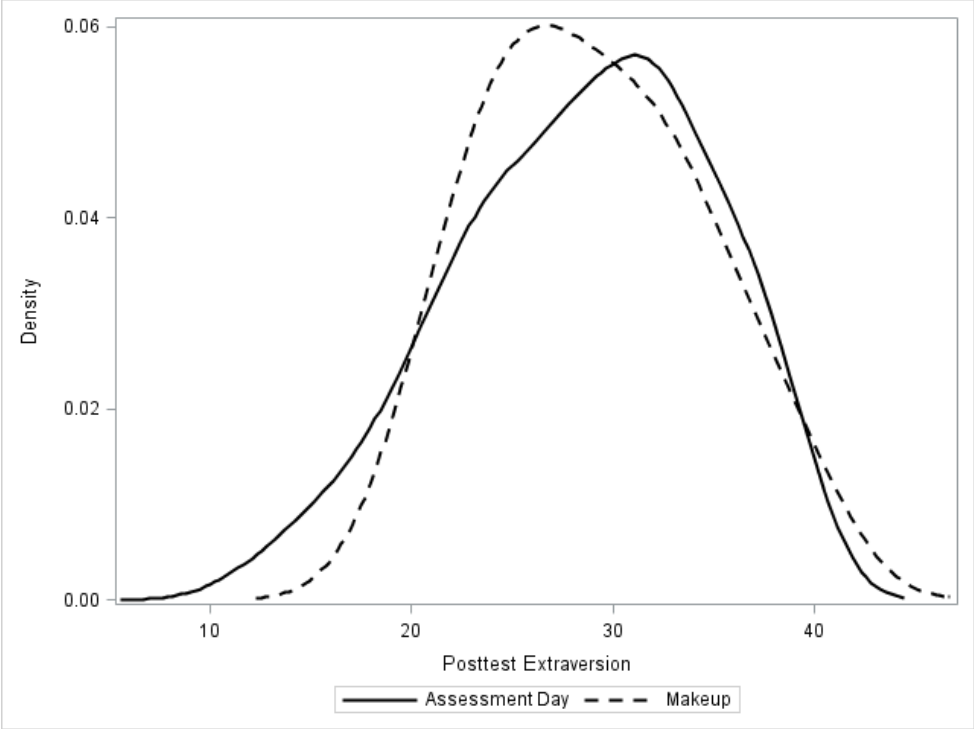
Posttest Openness



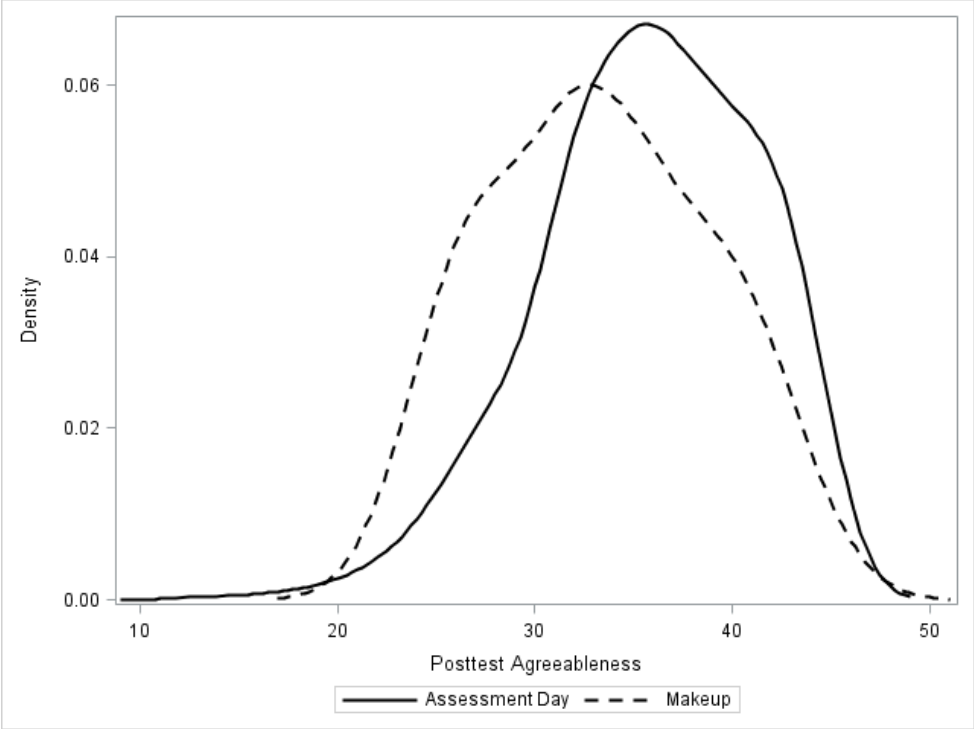
Posttest Conscientiousness



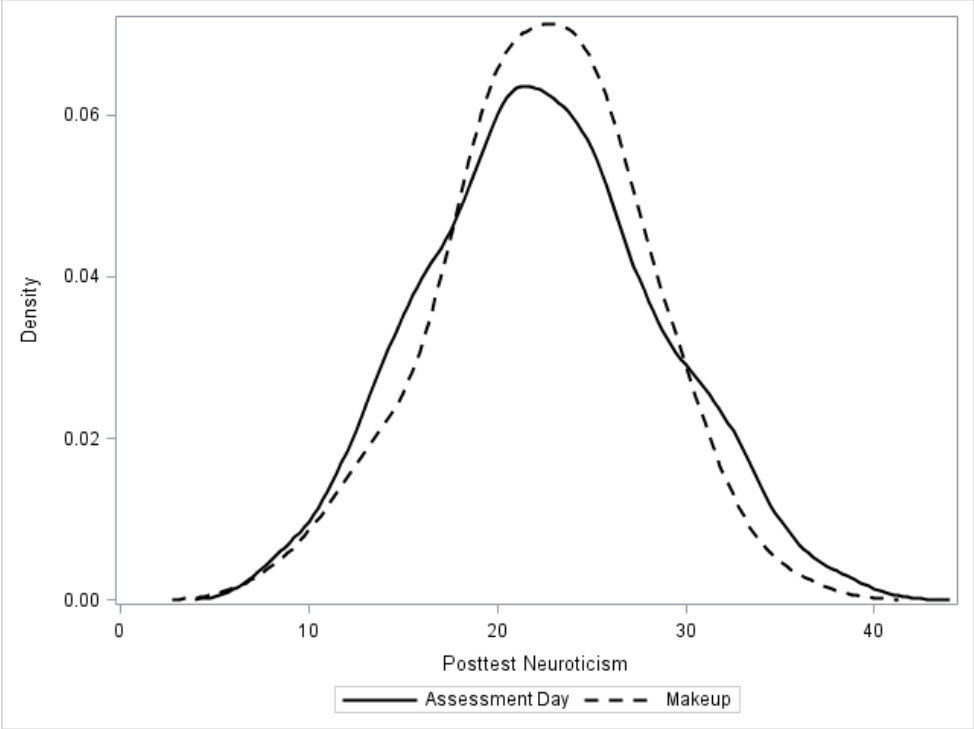
Posttest Extraversion



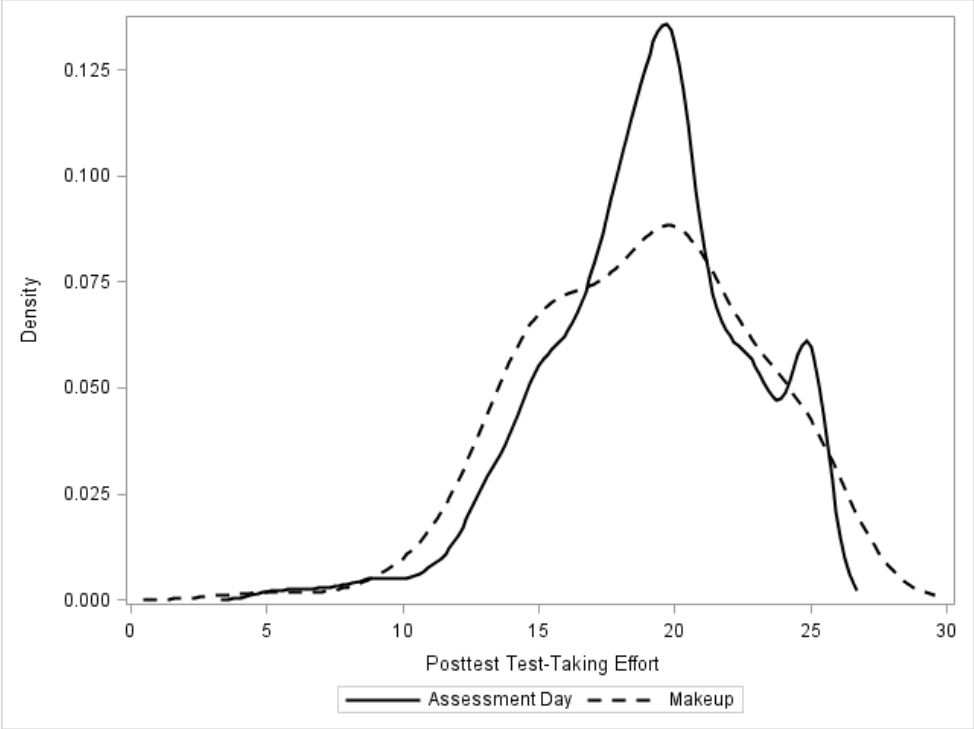
Posttest Agreeableness



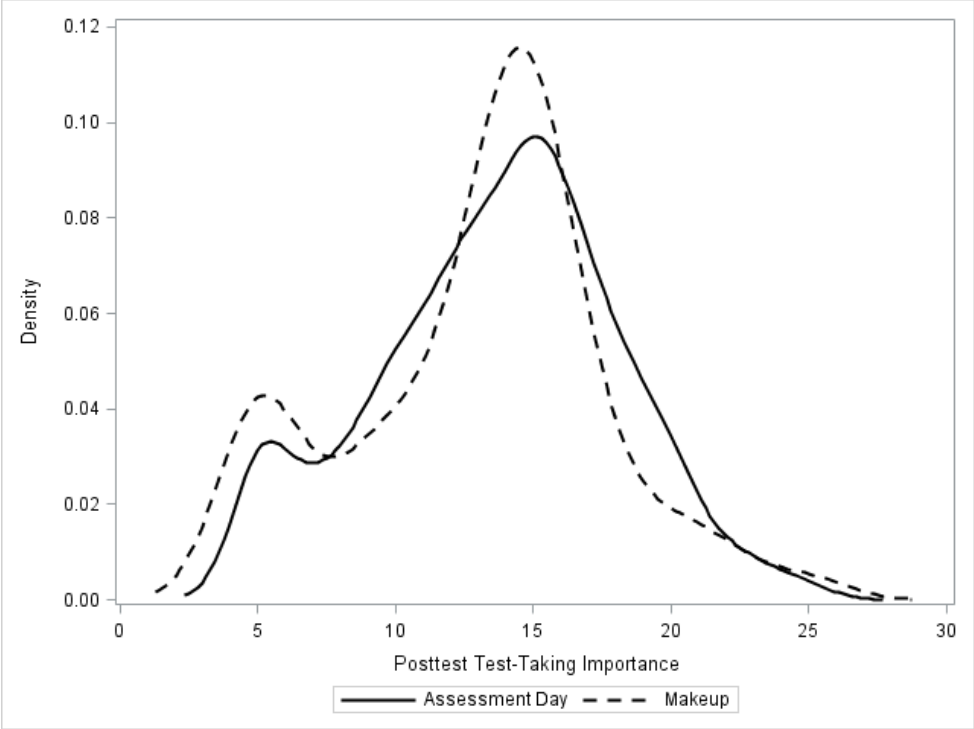
Posttest Neuroticism



Posttest Test-taking Effort



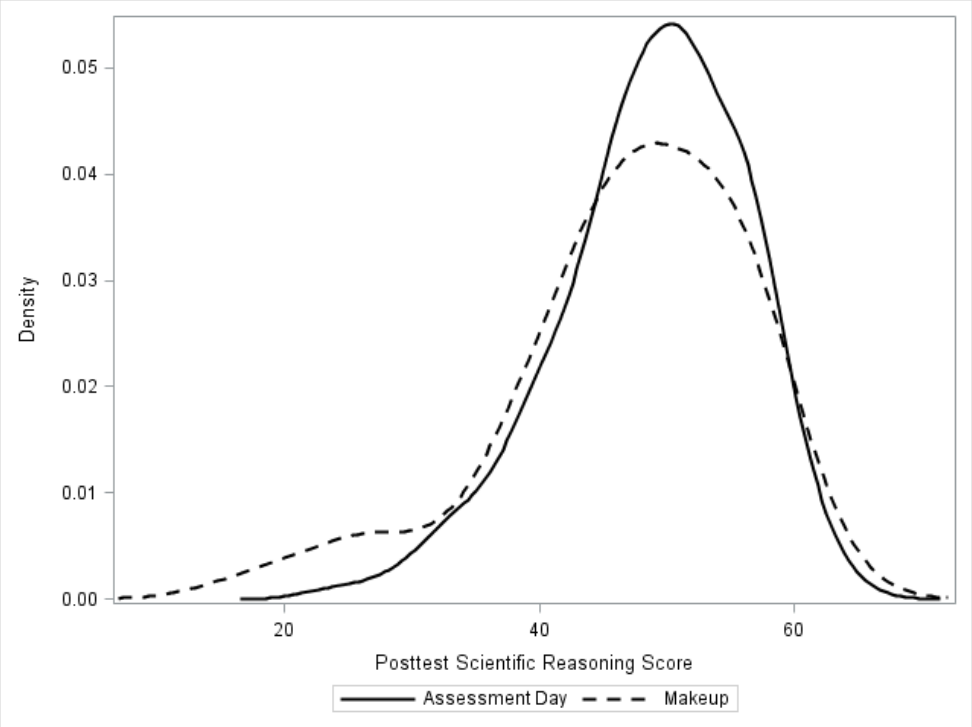
Posttest Test-taking Importance



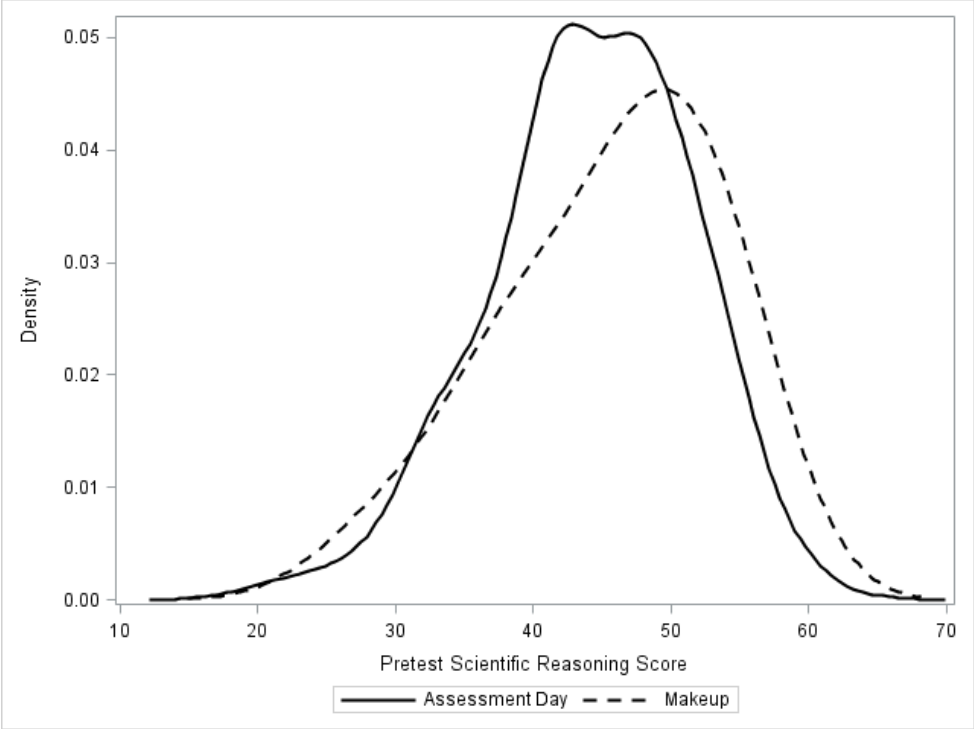
Appendix J

Histograms Comparing Assessment Day and Makeup Variable Distributions – Cognitive Sample

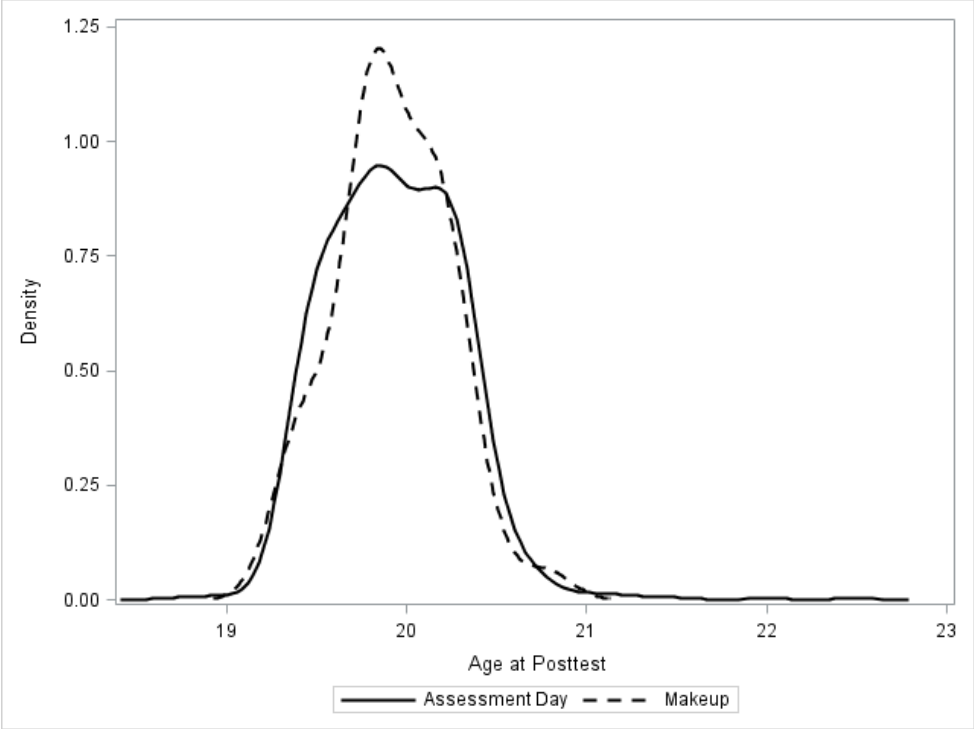
Posttest Scientific Reasoning



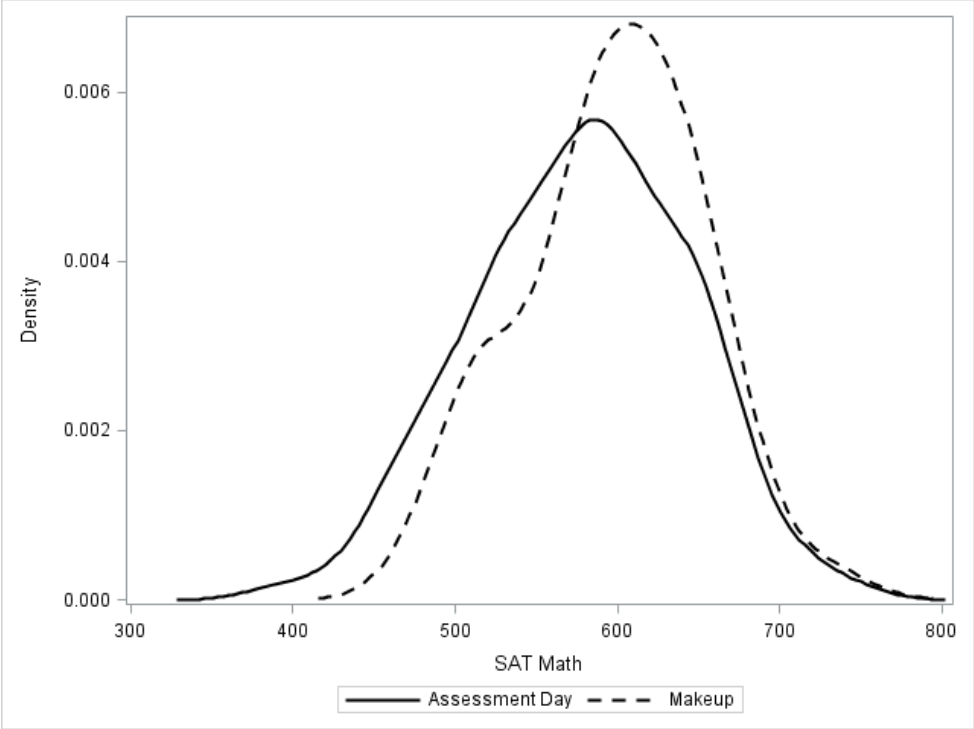
Pretest Scientific Reasoning



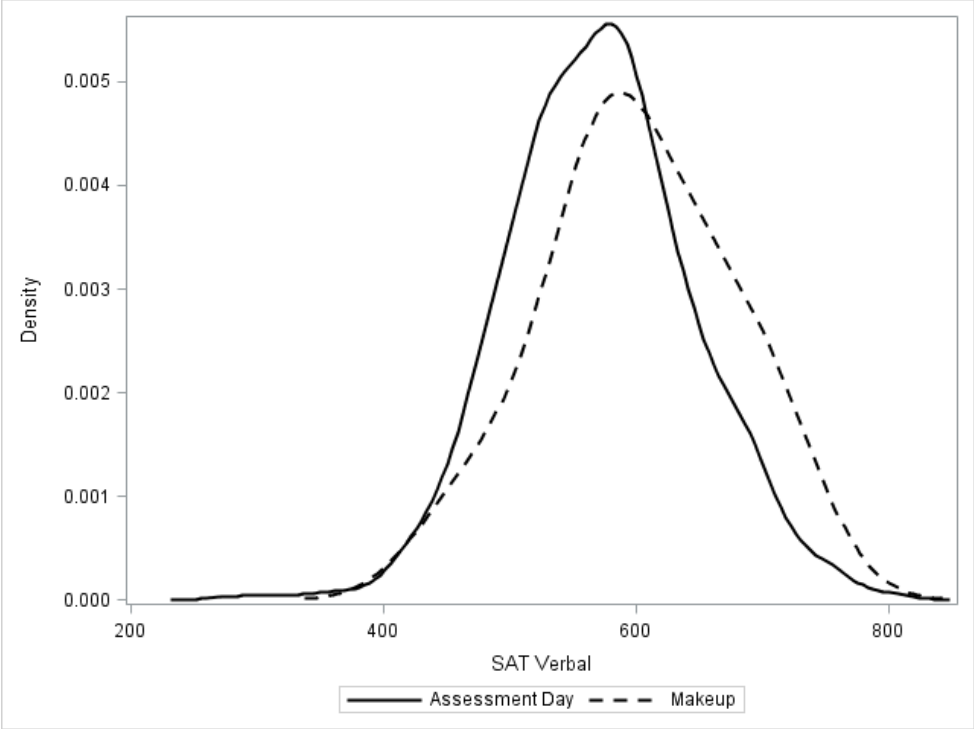
Age at Posttest



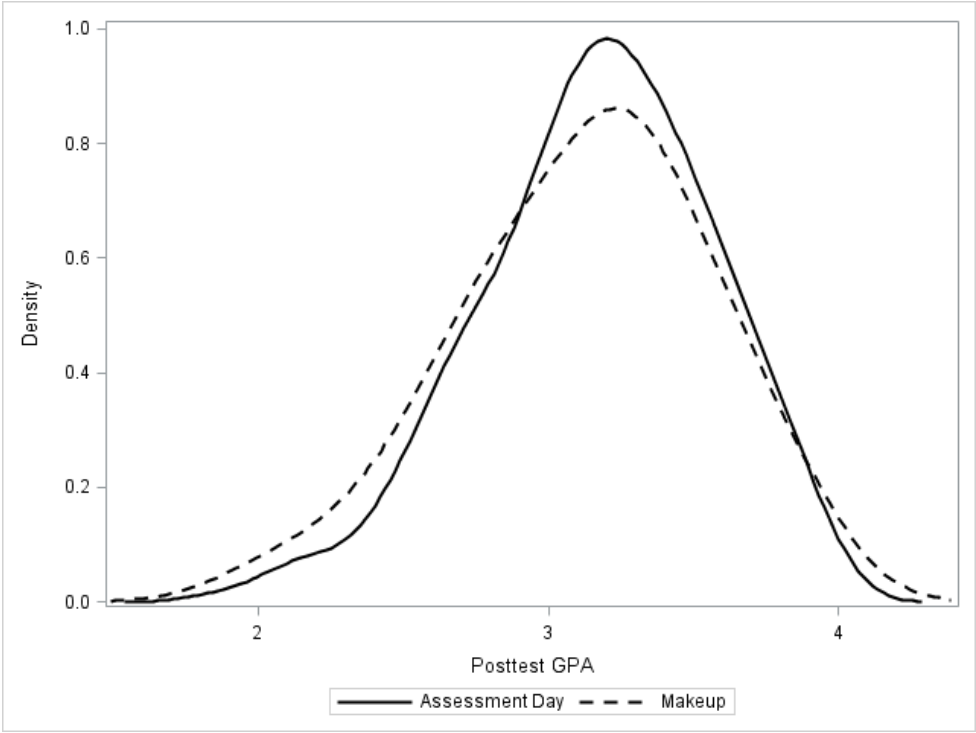
SAT Math



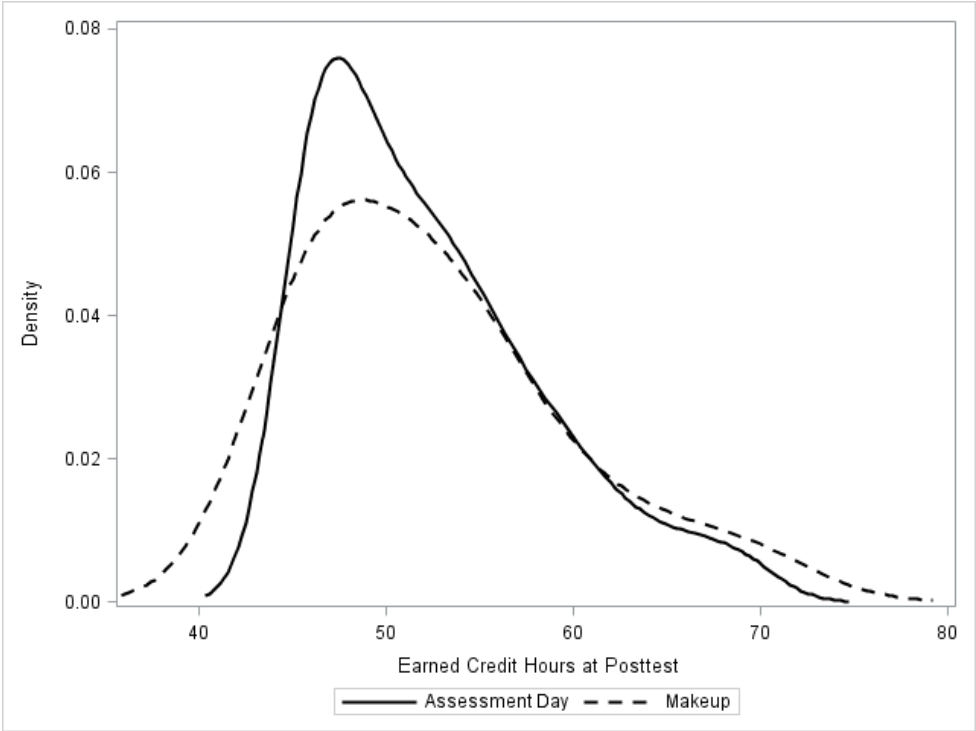
SAT Verbal



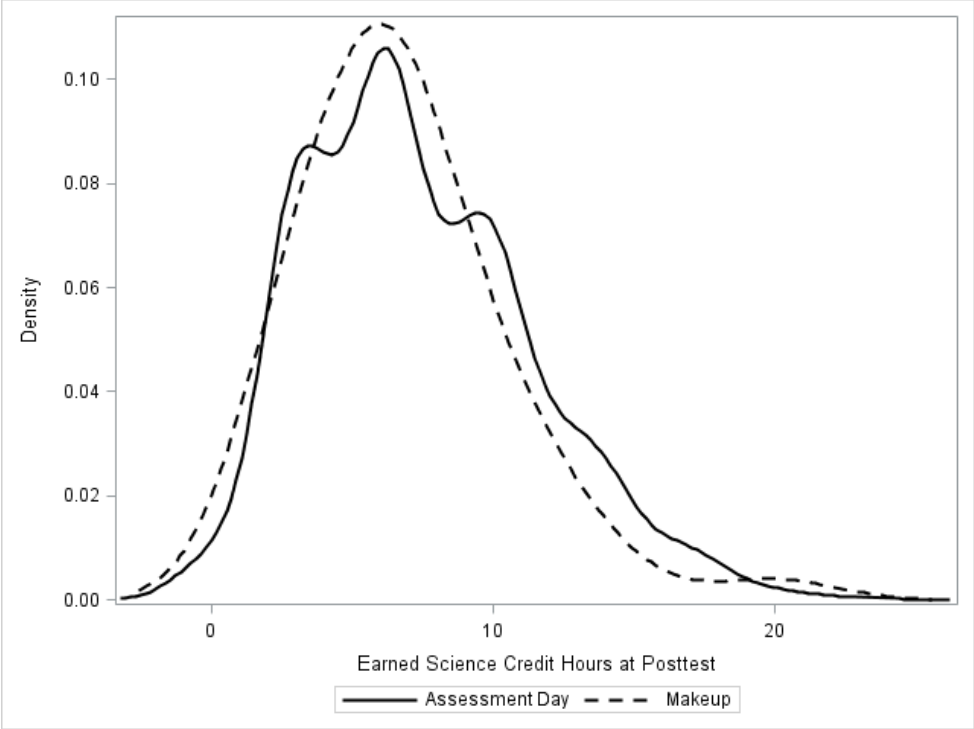
GPA



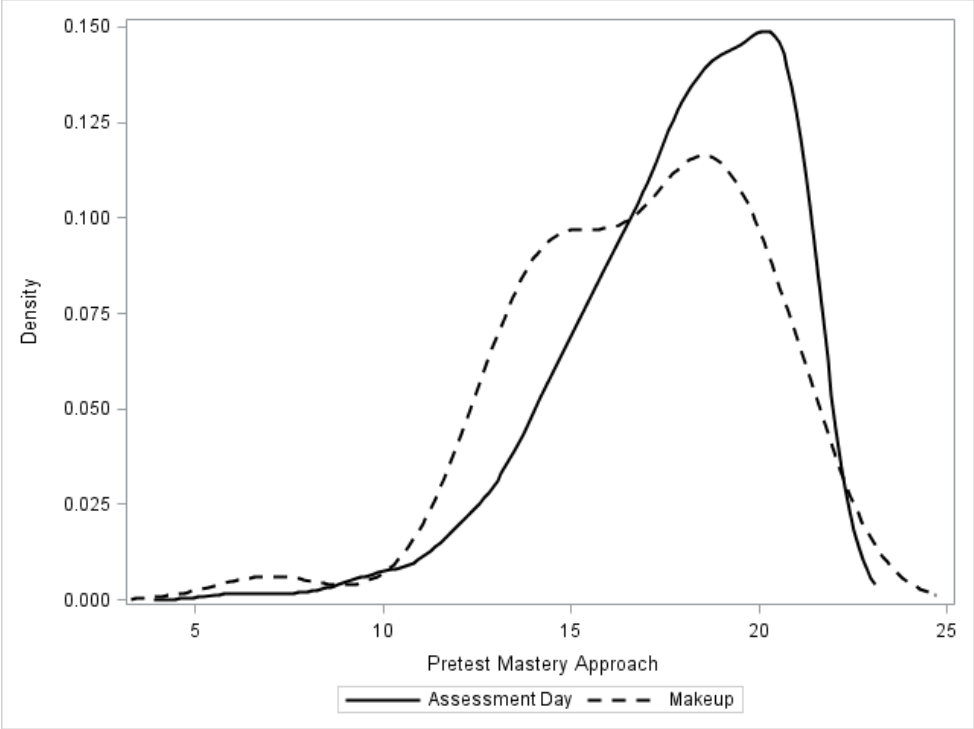
Credit Hours



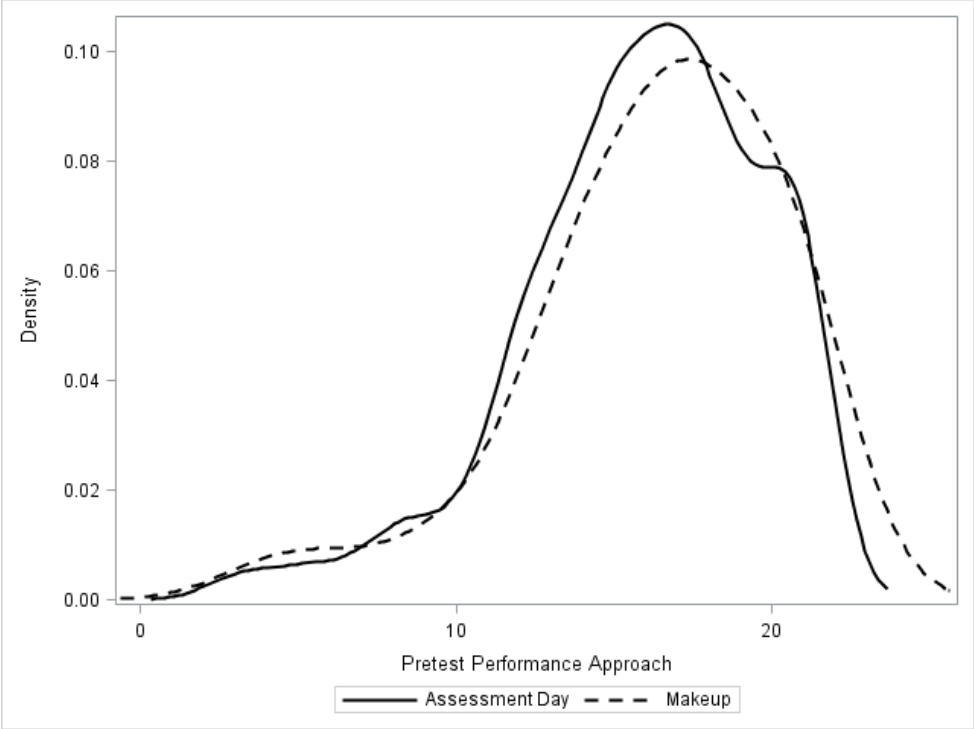
Science Credit Hours



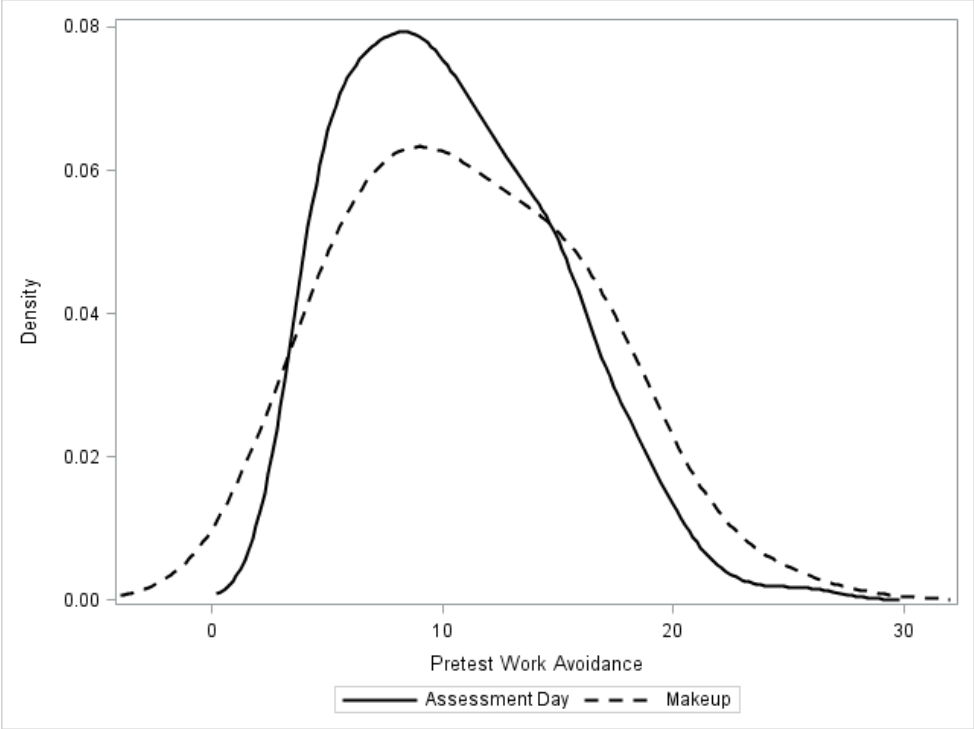
Pretest Mastery Approach



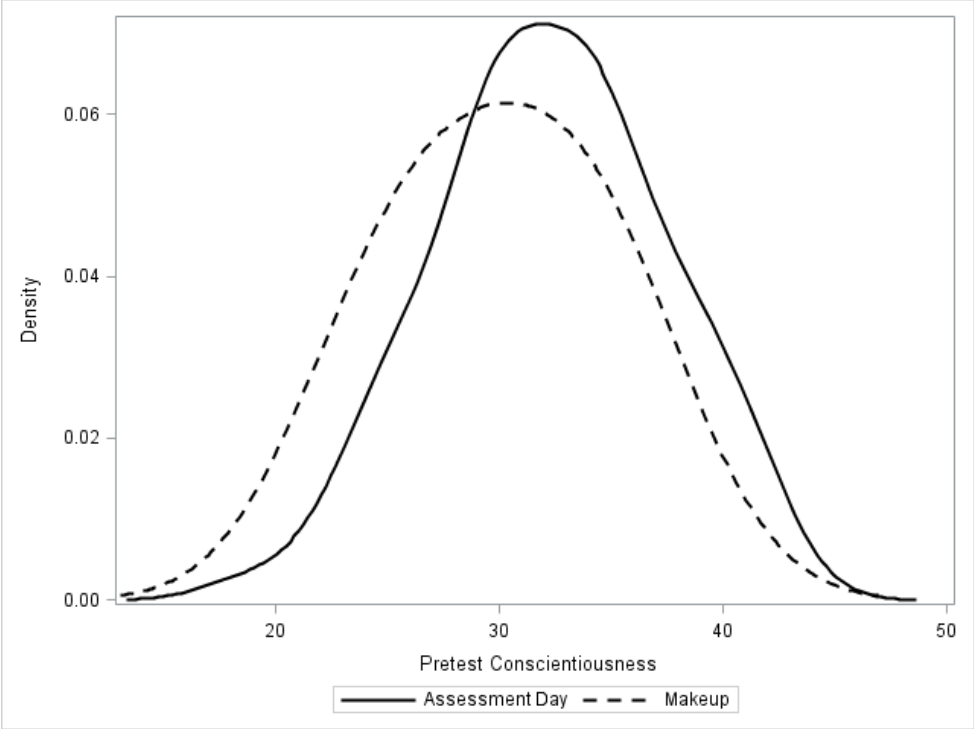
Pretest Performance Approach



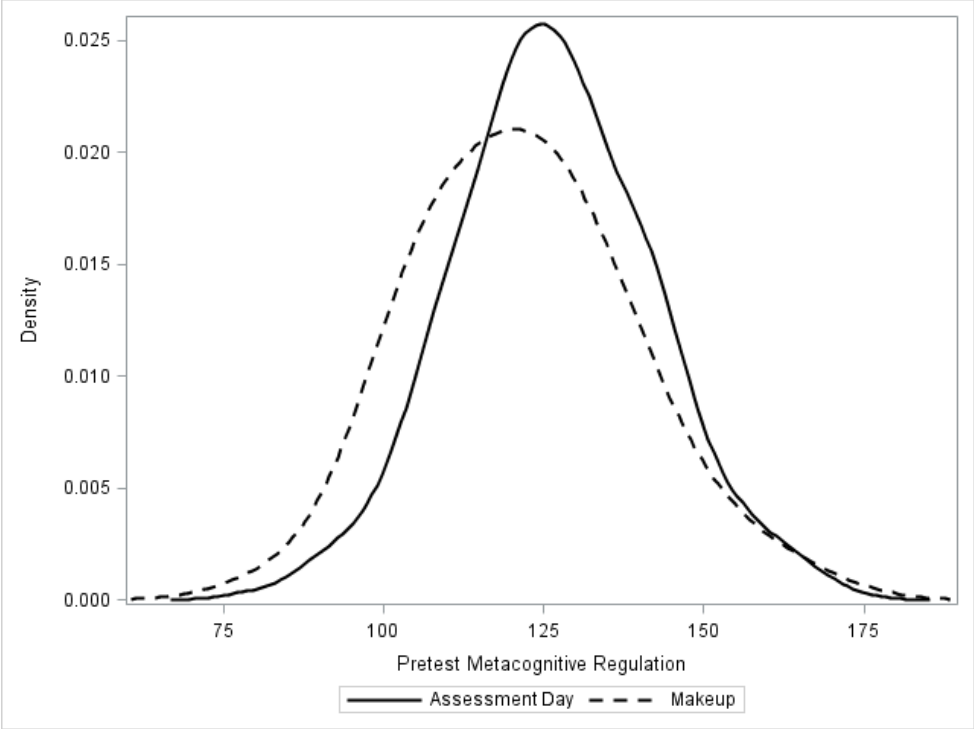
Pretest Work Avoidance



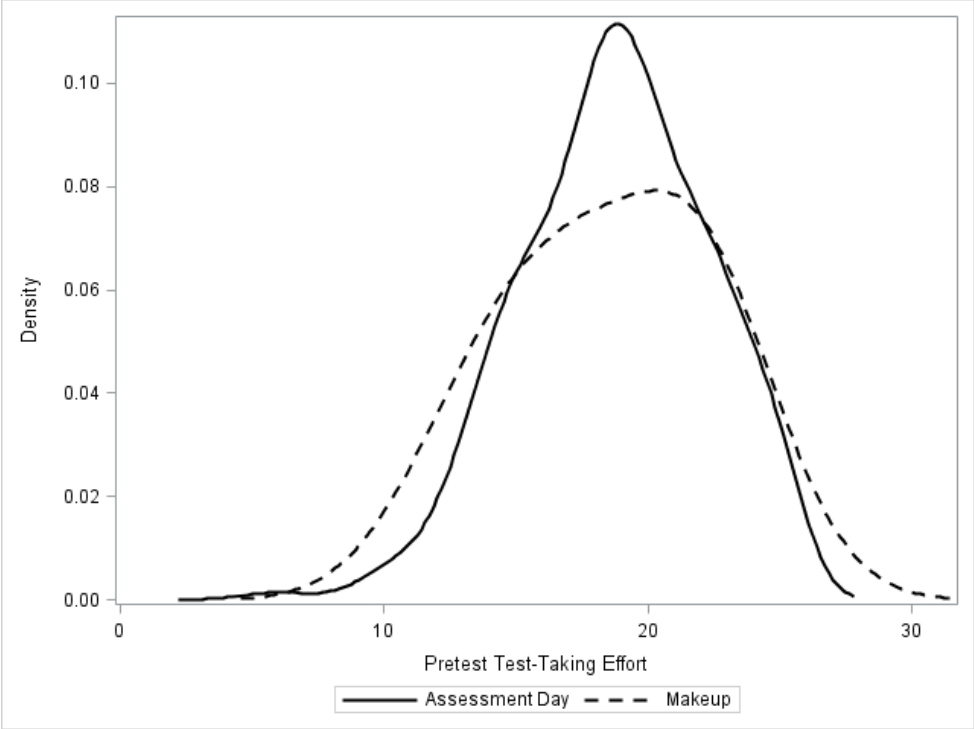
Pretest Conscientiousness



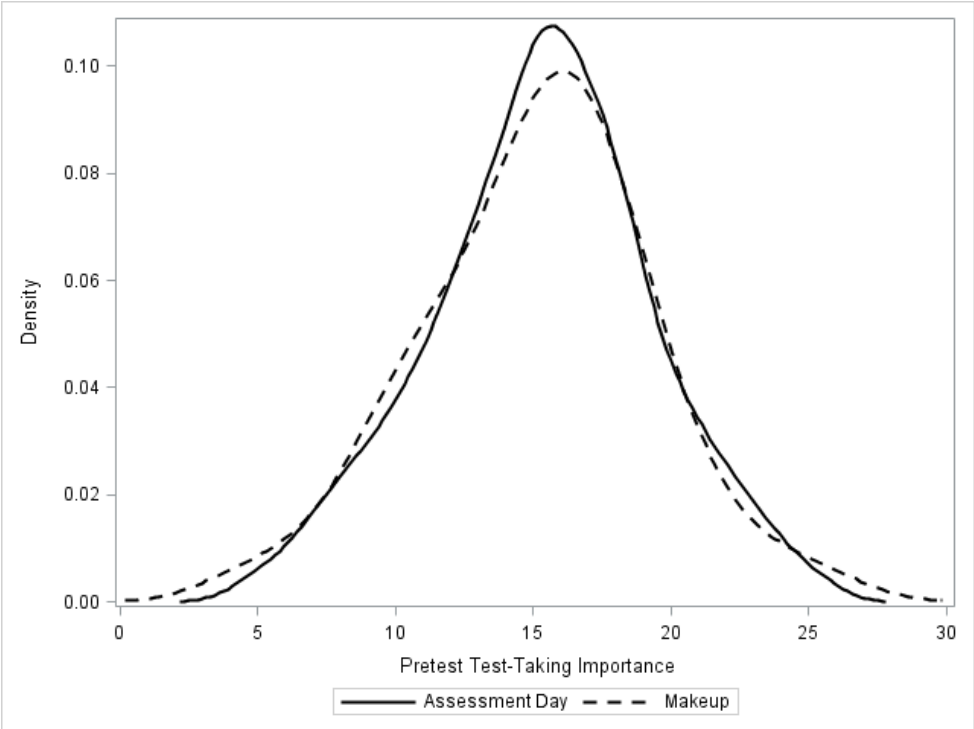
Pretest Metacognitive Regulation



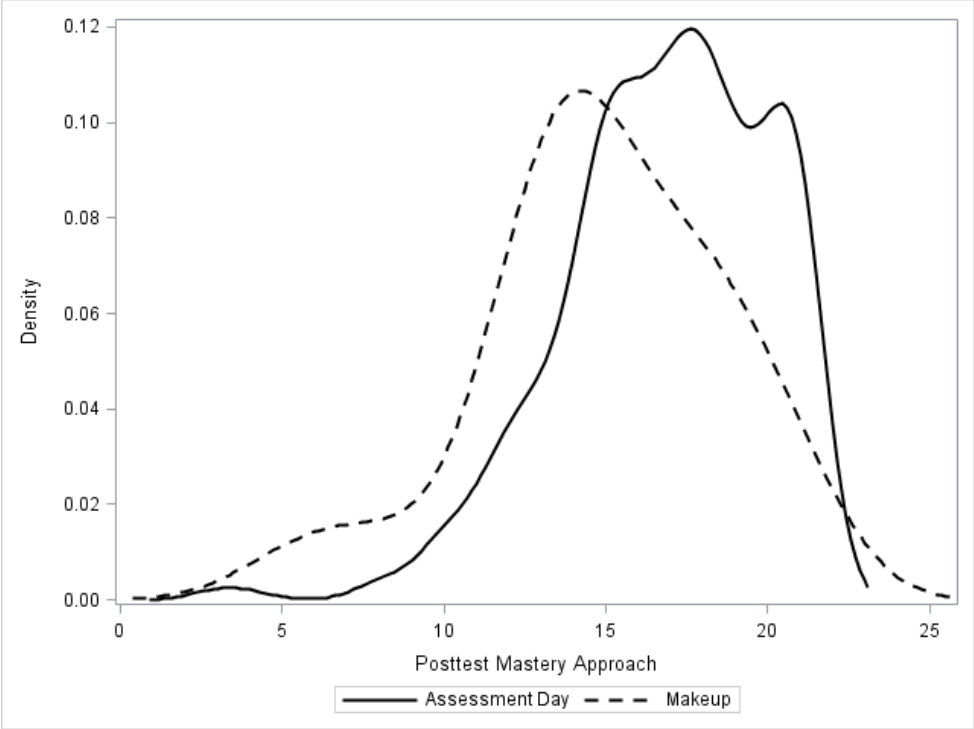
Pretest Test-taking Effort



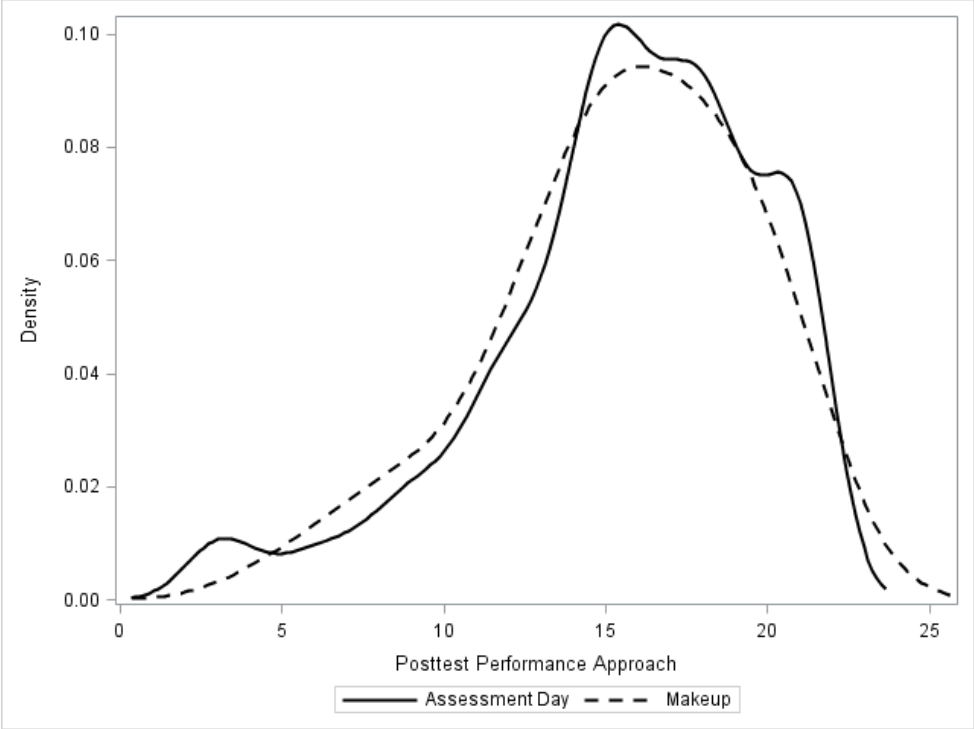
Pretest Test-taking Importance



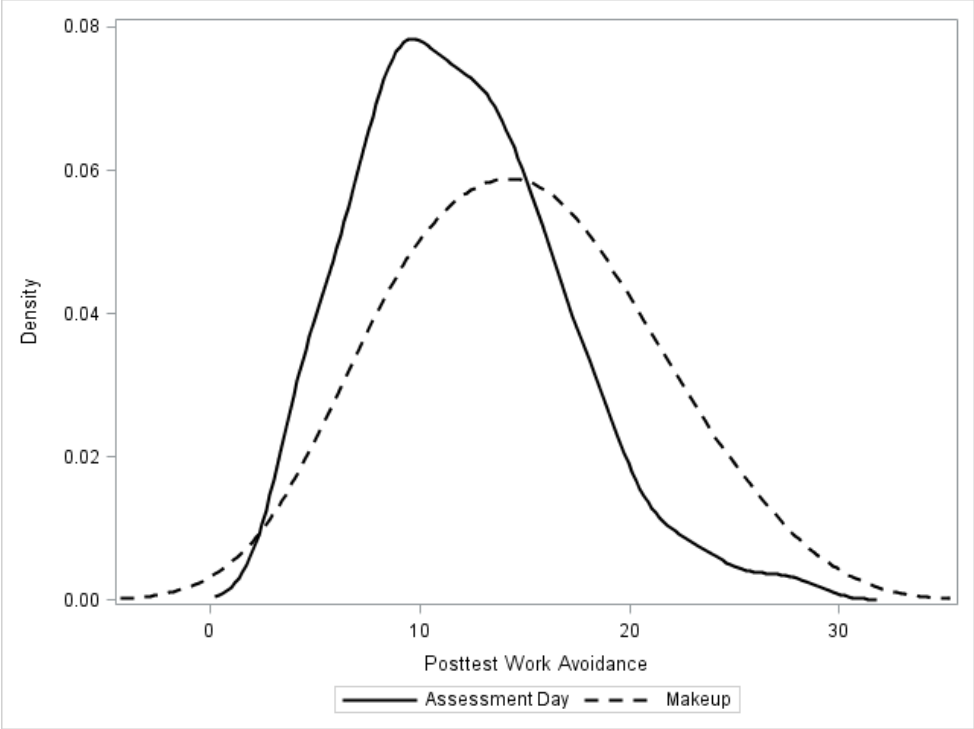
Posttest Mastery Approach



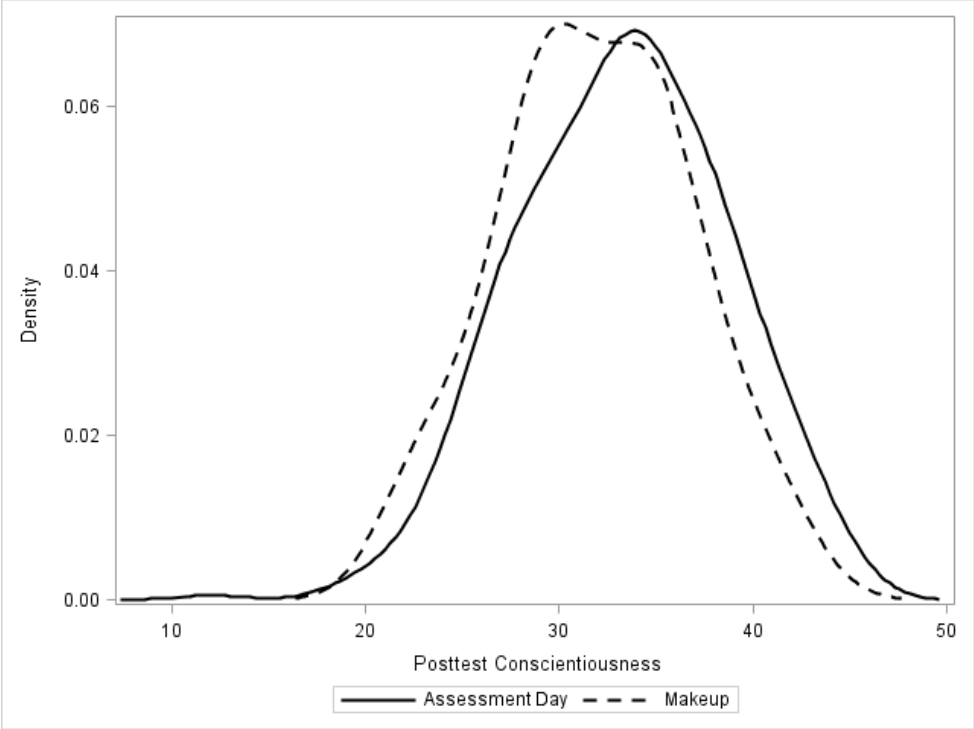
Posttest Performance Approach



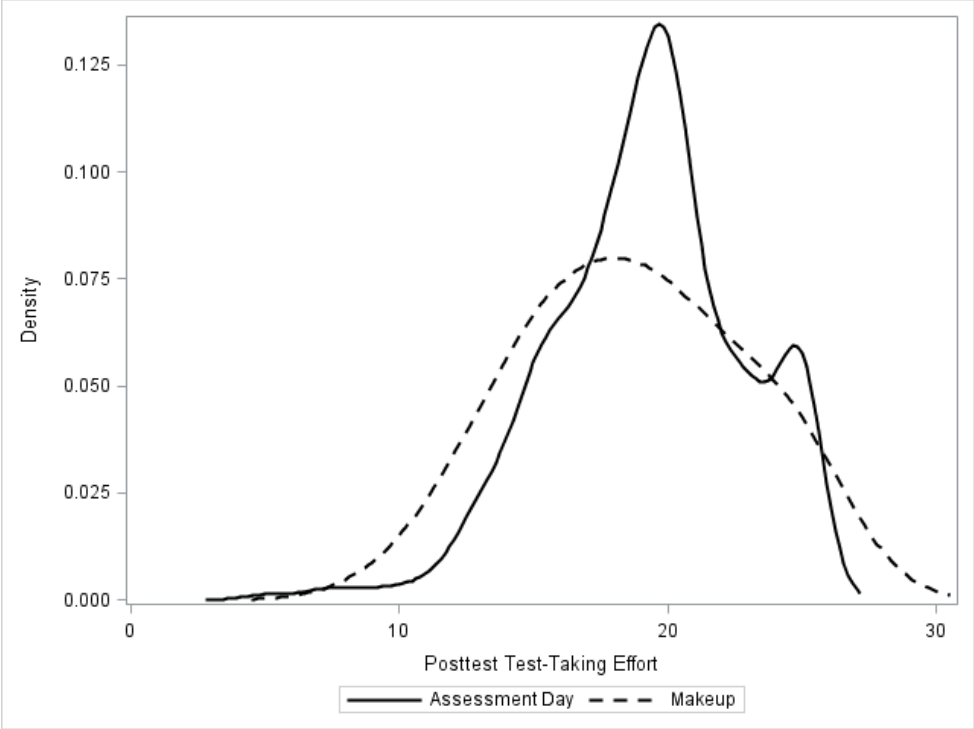
Posttest Work Avoidance



Posttest Conscientiousness



Posttest Test-taking Effort



Posttest Test-taking Importance